Seiberg-Witten Theory et Riemann Surfaces

par

Lois Sofia

Mémoire présenté au Département de mathématiques

en vue de l'obtention du grade de maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES

Université de Sherbrooke

Département de mathématiques

Sherbrooke, Québec, Canada.

Novembre 26, 2019

Le Novembre 29, 2019

le jury a accepté le mémoire de Madame Lois Sofia dans sa version finale.

Membres du jury

Professeur Vasilisa Shramchenko

Directrice de recherche

Département de mathématiques, Université de Sherbrooke

Professeur Patrick Labelle

Codirecteur de recherche

Département de mathématiques, Champlain College

Professeur Thomas Brüstle

Membre interne

Professeur Virginie Charette

Président-rapporteur

Département de mathématiques, Université de Sherbrooke

Sommaire

Dans cette thèse nous étudions les surfaces de Riemann dans le but de comprendre la théorie de Seiberg-Witten. Dans un article novateur, Seiberg et Witten ont obtenu l'approximation à basse énergie de la théorie de jauge de groupe SU(2) et ayant $\mathcal{N} = 2$ supersymétries en établissant un lien avec les propriétés d'une courbe elliptique. La théorie de jauge supersymétrique est complètement déterminée par une fonction holomorphe des champs, le prépotentiel, qui contient à la fois les corrections perturbatives et les corrections non-perturbatives. La correspondance à une courbe elliptique permit à Seiberg et Witten de calculer le prépotentiel de façon exacte, incluant les corrections non-perturbatives, qui sont extrêmement difficiles à obtenir par les méthodes traditionnelles.

Seiberg et Witten identifièrent la courbe elliptique en étudiant la structure des singularités de l'espace modulaire de la théorie et en comparant les monodromies du problème physique aux monodromies de la courbe elliptique. Dans cette thèse, nous passons en revue les arguments de Seiberg et Witten pour obtenir les monodromies du problème physique. Nous dérivons ensuite la même information à partir de la courbe elliptique en détail, exhibant ainsi la correspondance. Utilisant cette correspondance, nous dérivons de façon détaillée les expressions asymptotiques des variables (a_D, a) déterminées dans la théorie physique en calculant certaines intégrales de contour sur la courbe elliptique. Finalement, en utilisant des représentations des intégrales de contour en termes de fonctions hypergéométriques, nous reproduisons aussi le calcul des corrections non-perturbatives et déterminons quelques nombres, appelés nombres d'instantons, qui en découlent.

Abstract

In this thesis we study Riemann surfaces with a view to understanding Seiberg Written theory. In their seminal work, Seiberg and Witten derived the low energy approximation of the supersymmetric gauge theory having for gauge group SU(2) and $\mathcal{N} = 2$ supersymmetries by relating the problem to the data of an elliptic curve. The supersymmetric gauge theory is completely determined by a holomorphic function of the fields, the prepotential, that contains both perturbative and non-perturbative corrections of the theory. The identification with the elliptic curve allowed Seiberg and Witten to compute the prepotential of the theory exactly, including the non-perturbative corrections which are extremely difficult to obtain by traditional techniques.

Seiberg and Witten identified the elliptic curve by studying the singularity structure of the moduli space of the theory, and comparing the monodromy data of the physical problem with the monodromy data coming from the elliptic curve. In this thesis we review the arguments of Seiberg and Witten to arrive at the monodromy data of the physical problem. We then derive this same data from the elliptic curve in detail thus showing the correspondence. Using this correspondence, we derive the asymptotic expressions for the variables (a_D, a) determined in the physical theory by computing some contour integrals on the elliptic curve in some detail. Finally, using hypergeometric representations of the contour integrals, we also reproduce the calculation of the non-perturbative corrections and determine some of the so called instanton numbers which arise from them.

Acknowledgements

I would like to thank my advisors Professor Vasilisa Shramchenko and Professor Patrick Labelle for providing me academic guidance towards the thesis work and for their financial support during my studies. I would like to thank the members of the Department of Mathematics at Université de Sherbrooke for their support and hospitality.

Table of Contents

Chapter 1

Introduction

Quantum field theory, which describes the properties and interactions of elementary particles through the quantization of classical fields, has been one of the most fundamental advancements in the formulation of modern physics. Symmetry principles play a central role in the formulation of quantum field theories. The concept of gauge symmetry in quantum field theories, in particular, has proved to be a very fundamental idea in our understanding of nature. Three of the four fundamental forces of nature - electro-magnetism, strong and weak force - have a very successful formulation as gauge theories and have been very successful, for over 60 years, in explaining experimental observations, culminating with the discovery of the Higgs boson in 2012.

Supersymmetry, like gauge symmetry, is another symmetry principle that has been a very important part of the physics of the last thirty years. Supersymmetry is the symmetry that exchanges bosonic fields (i.e. quantum fields that commute) and fermionic fields (i.e. quantum fields that anticommute) with each other. Unlike gauge symmetry, however, the idea of supersymmetry has not been validated by experimental evidence. The search for the supersymmetric partners of known particles is one of the most active research areas in experimental particle physics.

Nevertheless, from a theoretical point of view, supersymmetric gauge theories have been one of the most important tools in our understanding of high energy physics. In recent times supersymmetric quantum fields theories, by themselves or within the framework of superstring theories, have become an important area of research in mathematical physics. Not only has pure mathematics been important to better understand supersymmetric quantum field theories, results from supersymmetric quantum field theories have led to major

insights in pure mathematics (Gromov-Witten theory, Mirror symmetry, knot theory, to name a few). Seiberg-Witten theory itself has led to study of invariants of compact smooth oriented 4 manifolds, Seiberg-Witten invariants. As such, even if supersymmetric quantum field theories turn out to not exist in nature, they will still remain a major tool in mathematical physics.

In this thesis we will consider a specific supersymmetric quantum field theory, the so-called $\mathcal{N} = 2$ supersymmetric SU(2) gauge theory (the meaning of these terms will be presented in chapter 3) and discuss Seiberg and Witten's solution [10] for the low energy effective theory. In quantum field theory, it is of particular interest to study the interactions of particles when they have small energies with respect to some scale (for example, the mass of the particles or other energy scale associated to the theory). The resulting approximation of the full theory (which is valid at all energies) is called a low energy effective quantum field theory, or effective theory for short. It is, however, extremely difficult to work out an explicit low energy effective theory.

The low energy effective theory of an $\mathcal{N} = 2$ supersymmetric theory can be determined by an exact calculation of the so-called $\mathcal{N} = 2$ prepotential. The prepotential receives both perturbative and non-perturbative contributions. Perturbative corrections can be calculated using the technology of Feynman diagrams and correspond to an expansion in powers of a small parameter called the gauge coupling constant. This expansion is equivalent to a Taylor expansion with each term in the expansion given by the sum of Feynman diagrams with fixed number of loops. The non-perturbative corrections are not analytic in the gauge coupling constant and cannot be obtained from Feynman diagrams. They are also organized as an expansion, but this time in powers of the inverse of the gauge coupling constant. In one interpretation, each term in this expansion corresponds to the interaction of an increasing number of soliton type excitations called "instantons". They require completely different techniques and are typically much more difficult to obtain than perturbative corrections. The contribution corresponding to the interaction of k instantons rapidly becomes unmanageable as k increases. It was therefore a major breakthrough when

Seiberg and Witten presented their solution for the $\mathcal{N} = 2$ supersymmetric SU(2) gauge theory.

The breakthrough was more conceptual than computational and therefore, more powerful, giving not only the low energy theory, but also a straightforward way of calculating the instanton corrections to any order. The idea of Seiberg and Witten was to identify the data associated to the low energy effective theory with an elliptic curve, with the parameter coming from the perturbative theory identified with one of the periods of a differential on the elliptic curve, and the gauge coupling identified with the modulus of the elliptic curve. What remained was the determination of the other period of the differential. They showed this to be given by a duality in the theory, with the other period of the differential determined by the dual parameter. Now, the gauge coupling, and from it the prepotential, including the non-perturbative contributions, could be determined exactly by calculating the period integrals of the elliptic curve.

Since then, a number of calculations of the first few instanton terms have been obtained using more standard - and much more difficult - approaches and the results of Seiberg and Witten have been confirmed.

The work of Seiberg-Witten represented a major breakthrough in our understanding of supersymmetric theories and have since been extended to many other classes of supersymmetric theories. In this thesis we will present the mathematical details of their work, providing only the minimum physical input to understand their discovery.

The organization of the thesis is as follows. In chapter 2 we introduce Riemann surfaces from a topological and algebraic point of view. In chapter 3, we introduce Seiberg-Witten theory, and explain how the data of an elliptic curve emerges from the low energy effective theory. In chapter 4 we solve the theory by calculating the period integrals and the instanton numbers.

Chapter 2

Riemann Surfaces

2.1 Introduction

In this chapter we introduce and study Riemann surfaces with a view to understanding the geometry of the Seiberg-Witten curve. We collect some important facts and results which make the thesis as self-contained as possible to discuss the main idea of the thesis – Solution of the Seiberg-Witten gauge theory. The chapter follows the references [1], [2], [3], [4], [5], [6], [7], [8] [9].

2.2 Riemann Surfaces

A *Riemann surface* is a complex one-dimensional (real two-dimensional) analytic manifold with a complex structure on it. Let us elaborate on this notion one definition at a time. A surface, $\mathcal{R}$, is a two-dimensional real topological manifold. As a manifold we can cover it with open sets $\{\mathcal{U}_\alpha\}_{\alpha\in A}$ for some counting set A i.e. $\cup_{\alpha\in A}\mathcal{U}_\alpha = \mathcal{R}$ along with homeomorphisms identifying the open sets with open subsets of $\mathbb{R}^2$ i.e. $\phi_\alpha : \mathcal{U}_\alpha \to V_\alpha \in \mathbb{R}^2$. The pair $(\mathcal{U}_\alpha, \phi_\alpha)$ is called a chart while the whole family $\{(\mathcal{U}_\alpha, \phi_\alpha)\}$ is known as an *atlas*. Let $\phi : \mathcal{U} \to V$ be a chart on $\mathcal{R}$, with $p \in U$, then $z = \phi(x)$ for x in the neighborhood of p is called the local coordinate on $\mathcal{U}$.

Of course, one can find many different possible atlases on $\mathcal{R}$ by covering it with different charts. Given two different charts $(\mathcal{U}_\alpha, \phi_\alpha)$ and $(\mathcal{U}_\beta, \phi_\beta)$, one requires a compatibility between them where they overlap:

$$\phi_{\alpha\beta} = \phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(\mathcal{U}_\alpha \cap \mathcal{U}_\beta) \to \phi_\beta(\mathcal{U}_\alpha \cap \mathcal{U}_\beta) \tag{2.1}$$

for functions $\phi_{\alpha\beta}$. The functions $\phi_{\alpha\beta}$ in (2.1) are the transition functions for going from one chart to another. We could require different properties for the transition functions to emphasize the structure we are studying. For our case, we identify $\mathbb{R}^2$ with $\mathbb{C}$, and require the transition functions $\phi_{\alpha\beta}$ to be holomorphic functions making the surface $\mathcal{R}$ a complex manifold. Two atlases, $\{(\mathcal{U}_\alpha, \phi_\alpha)\}$ and $\{(\widetilde{\mathcal{U}}_\alpha, \widetilde{\phi}_\alpha)\}$ on $\mathcal{R}$ are compatible if $\{(\mathcal{U}_\alpha, \phi_\alpha)\} \cup \{(\widetilde{\mathcal{U}}_\alpha, \widetilde{\phi}_\alpha)\}$ is again a complex atlas. This defines an equivalence relation on the set of holomorphic atlases. An equivalence class of complex atlases is called a complex structure on $\mathcal{R}$. Putting all these ideas back together we get a Riemann surface as a complex one-dimensional manifold with a choice of complex structure on it.

Examples of Riemann surfaces:

Complex plane

The simplest example of a Riemann surface is the complex plane itself. We need just one chart, $\mathcal{U} = \mathbb{C}$, to cover it and the associated homeomorphism is the identity mapping : $id : \mathbb{C} \to \mathbb{C}$.

Open sets of $\mathbb{C}$

Every open set of the complex plane is again a Riemann surface. In particular

- The unit disk $\mathbb{D} = \{z \in \mathbb{C} \mid |z| < 1\}$
- The upper half plane : $\{\mathbb{H} = \{z \in \mathbb{C} \mid \mathrm{Im} z > 0\}$
- The complex plane with the origin removed $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$

are all important examples of Riemann surfaces. In general every open subset of a Riemann surface is again a Riemann surface.

Compact Riemann surfaces

Compact oriented (two dimensional) surfaces are spheres with handles and are classified up to homeomorphism by the number of handles on them$(Fig)^*$. The number of handles is a topological invariant known as the genus of the surface. We study examples of compact Riemann surfaces in genus 0 and genus 1 below.

Genus zero – The sphere

A genus zero surface is one without any handles – a sphere – and every genus zero surface is homeomorphic to the sphere. There is more than one useful way of realizing the genus zero surface and we see two equivalent realizations below. Both the realizations are homeomorphic, as topological manifolds, to the two-sphere, S^2.

Riemann sphere

The Riemann sphere (or the extended complex plane) is the complex plane together with the point $\{\infty\}$, called the "point at infinity" : $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. It cannot be covered by a single chart like $\mathbb{C}$ since at $z \to \infty$ the coordinate is ill defined. To cover the point at infinity, we need an additional chart which would be well defined at $\{\infty\}$. Thus the atlas of $\widehat{\mathbb{C}}$ consists of two charts given by:

$$\begin{aligned} &\mathcal{U}_0 := \widehat{\mathbb{C}} \setminus \{\infty\}, \quad \phi_0 : \mathcal{U}_0 \to \mathbb{C}, \quad \phi_0(z) = z \ \text{ and} \\ &\mathcal{U}_\infty := \widehat{\mathbb{C}} \setminus \{0\}, \quad \phi_\infty : \mathcal{U}_\infty \to \mathbb{C}, \quad \phi_\infty(z) = 1/z. \end{aligned} \tag{2.2}$$

There are two transition functions

$$\phi_{0\infty} = \phi_\infty \circ \phi_0^{-1}, \quad \phi_{\infty 0} = \phi_0 \circ \phi_\infty^{-1} \quad : \quad \mathbb{C}^* \to \mathbb{C}^* \tag{2.3}$$

which are both holomorphic

$$\phi_{0\infty}(z) = \phi_{\infty 0}(z) = 1/z \ . \tag{2.4}$$

The complex projective line – $\mathbb{CP}^1$

The complex projective line is defined as the set of all ordered pairs of complex numbers $\{(z_0, z_1) \in \mathbb{C}^2 \mid (z_0, z_1) \neq (0,0)\}$ where we identify pairs up to scalar multiplication, i.e.

$$\mathbb{CP}^1 = \{\mathbb{C}^2 \setminus \{(0,0)\}\}/\sim \tag{2.5}$$

where, denoting $z = (z_0, z_1)$ and $w = (w_0, w_1)$, the equivalence relation is given by

$$z \sim w \text{ if } \; z = \lambda w, \;\; \text{for some } \lambda \in \mathbb{C}^* \; . \tag{2.6}$$

The coordinates $z = (z_0, z_1)$ are called homogeneous coordinates of $\mathbb{CP}^1$ denoted by $[z_0, z_1]$. The point at infinity corresponds to $[1, 0]$. The complex projective line is covered by an atlas consisting of two charts given by:

$$\begin{aligned} &\mathcal{U}_0 : \{[z_0, z_1] \in \mathbb{CP}^1 \mid z_0 \neq 0\}, \quad \phi_0 : \mathcal{U}_0 \to \mathbb{C}, \quad [z_0, z_1] \mapsto z_1/z_0, \text{ and} \\ &\mathcal{U}_1 : \{[z_0, z_1] \in \mathbb{CP}^1 \mid z_1 \neq 0\}, \quad \phi_1 : \mathcal{U}_1 \to \mathbb{C}, \quad [z_0, z_1] \mapsto z_0/z_1. \end{aligned} \tag{2.7}$$

Genus one – Complex Tori

Tori are examples of a genus 1 Riemann surfaces. Topologicaly, a torus is a surface shaped like a donut. To describe it as a complex manifold with a complex structure on it, let ω_1, ω_2 be two complex numbers which are linearly independent over $\mathbb{R}$. Without loss of generality we may take $\mathrm{Im}(\omega_2/\omega_1) > 0$. The two complex numbers generate a lattice

$$L(\omega_1, \omega_2) \equiv \{m\omega_1 + n\omega_2, \mid n, m \in \mathbb{Z}\} \; . \tag{2.8}$$

The manifold $\mathbb{C}/L(\omega_1, \omega_2)$ is obtained by identifying the points $z_1, z_2 \in \mathbb{C}$ if $z_1 - z_2 = m\omega_1 + n\omega_2$ for $m, n \in \mathbb{Z}$. Topologically, $\mathbb{C}/L(\omega_1, \omega_2)$ is a parallelogram with vertices $(0, \omega_1, \omega_2, \omega_1 + \omega_2)$ with the opposite sides identified. This surface is homeomorphic to a torus T^2, which is a genus 1 surface, as shown in the figure. The manifold $\mathbb{C}/L(\omega_1, \omega_2)$ has a naturally induced complex structure from $\mathbb{C}$. Thus, the pairs (ω_1, ω_2) describe the complex structures on the torus. The first obvious question that arises is if the pair (ω_1, ω_2) is unique

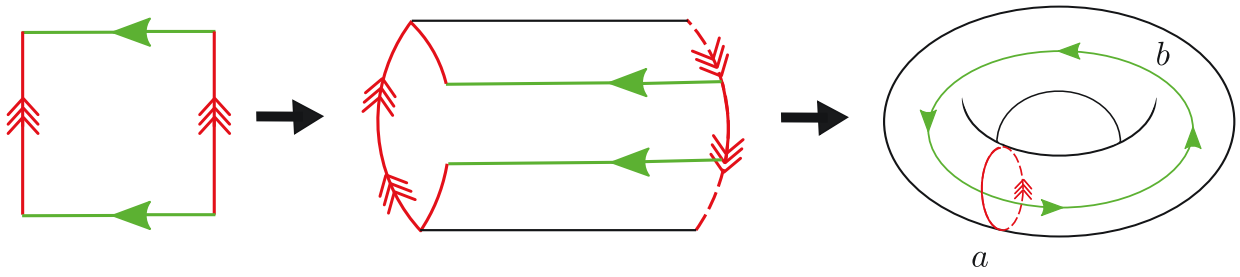

Figure 2.1: The Torus and a and b cycles

in describing the given complex structure, or if more than one pair could correspond to the same complex structure. A quick way to see the answer is to consider the pair $(-\omega_2, \omega_1)$. This would generate the exact same parallelogram in the complex plane and lead to the same torus T^2. Thus, intuitively, it should generate the same complex structure. However, there is no clear way to see that the transformation $(\omega_1, \omega_2) \mapsto (-\omega_2, \omega_1)$ is trivial. Thus, there should be more than one set of parameters that correspond to the same complex structure. It turns out that two pairs (ω_1, ω_2), $\mathrm{Im}\frac{\omega_2}{\omega_1} > 0$, and (ω_1', ω_2'), $\mathrm{Im}\frac{\omega_2'}{\omega_1'} > 0$ define the same complex structure if and only if they are related by a $PSL(2,\mathbb{Z})$ matrix, i.e. $(\omega_1, \omega_2) \sim (\omega_1', \omega_2')$ if and only if there exists a matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in PSL(2,\mathbb{Z}) = SL(2,\mathbb{Z})/\{I, -I\} \tag{2.9}$$

such that

$$\begin{pmatrix} \omega_1' \\ \omega_2' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}. \tag{2.10}$$

Here $\mathrm{SL}(2,\mathbb{Z})$ is the set of all 2×2 matrices with integer coefficients and unit determinant, and the matrices M and $-M$ are identified in $\mathrm{PSL}(2,\mathbb{Z})$. The equivalence $(\omega_1, \omega_2) \sim (\omega_1', \omega_2')$ can be seen as follows. Let

$$\begin{pmatrix} \omega_1' \\ \omega_2' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} \qquad \text{for} \qquad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z}) \,. \tag{2.11}$$

Since $\omega_1', \omega_2' \in L(\omega_1, \omega_2)$, we have that $L(\omega_1', \omega_2') \subset L(\omega_1, \omega_2)$. Inverting (2.11), we get

$$\begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \begin{pmatrix} \omega_1' \\ \omega_2' \end{pmatrix} \tag{2.12}$$

and so we get that $L(\omega_1, \omega_2) \subset L(\omega_1', \omega_2')$. Thus, $L(\omega_1, \omega_2) = L(\omega_1', \omega_2')$. Conversely, if we have that $L(\omega_1, \omega_2) = L(\omega_1', \omega_2')$, then ω_1' and ω_2' are lattice points of $L(\omega_1, \omega_2)$ and hence can be expressed as

$$\omega_1' = d\,\omega_1 + c\,\omega_2, \quad \omega_2' = b\,\omega_1 + a\,\omega_2, \qquad \text{for some } a, b, c, d \in \mathbb{Z}. \tag{2.13}$$

Similarly, ω_1 and ω_2 can be expressed as

$$\omega_1 = d'\,\omega_1' + c'\omega_2', \quad \omega_2 = b'\,\omega_1' + a'\omega_2' \qquad \text{for some } a', b', c', d' \in \mathbb{Z}. \tag{2.14}$$

From (2.13) and (2.14) we have

$$\begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} = \begin{pmatrix} d' & c' \\ b' & a' \end{pmatrix} \begin{pmatrix} \omega_1' \\ \omega_2' \end{pmatrix} = \begin{pmatrix} d' & c' \\ b' & a' \end{pmatrix} \begin{pmatrix} d & c \\ b & a \end{pmatrix} \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix} \tag{2.15}$$

which implies

$$\begin{pmatrix} d' & c' \\ b' & a' \end{pmatrix} \begin{pmatrix} d & c \\ b & a \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{2.16}$$

Taking determinants on both sides of the above equations we get $(a'd' - b'c')(ad - bc) = 1$. Since a', b', c', d' and a, b, c, d are all integers, this is possible only if $ad - bc = \pm 1$. Further, the condition $\mathrm{Im}\,(\omega_2/\omega_1) > 0$ gives

$$\mathrm{Im}\left(\frac{\omega_2'}{\omega_1'}\right) = \mathrm{Im}\left(\frac{b\omega_1 + a\omega_2}{d\omega_1 + c\omega_2}\right) = \frac{ad - bc}{|c(\omega_2/\omega_1) + d|^2}\ \mathrm{Im}\left(\frac{\omega_2}{\omega_1}\right) > 0, \tag{2.17}$$

from which we must have that $ad - bc > 0$. Thus,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2, \mathbb{Z}). \tag{2.18}$$

It is also straightforward to note that the matrices A and $-A$ in $\mathrm{SL}(2, \mathbb{Z})$ define the same lattice and hence two lattices that are related by $\mathrm{PSL}(2, \mathbb{Z}) = \mathrm{SL}(2, \mathbb{Z})/\{I, -I\}$ are the same.

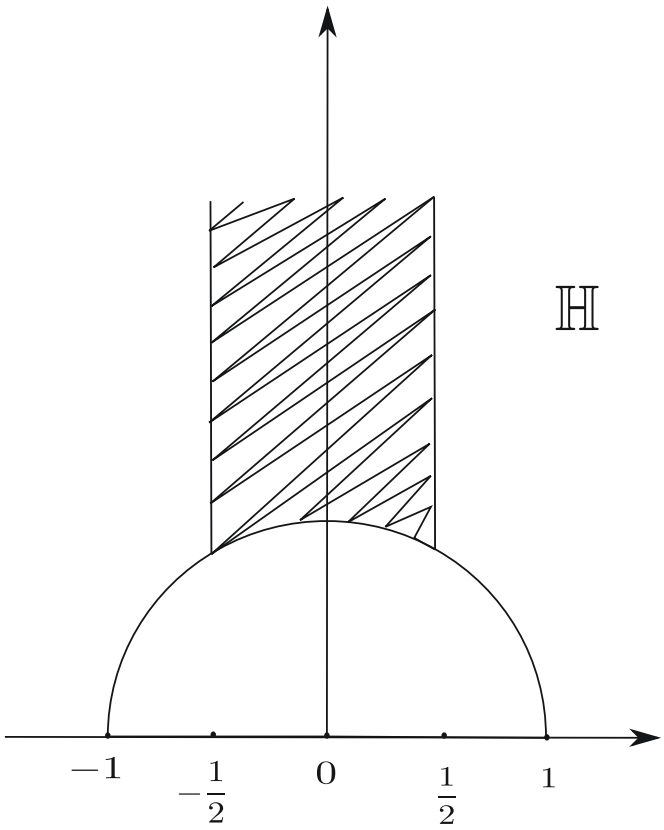

Figure 2.2: The quotient space $\mathbb{H}/PSL(2,\mathbb{Z})$

Define the quantity

$$\tau = \omega_2/\omega_1 \in \mathbb{H} \equiv \{z \in \mathbb{C} \mid \text{ Im } z > 0\} \tag{2.19}$$

to specify the complex structure of T^2. Without loss of generality, we can take 1 and τ to be the generators of a lattice. From our previous discussion we have that, τ and $\tau' = (a\tau + b)/(c\tau + d)$ define the same complex structure if

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \text{PSL}(2,\mathbb{Z}). \tag{2.20}$$

The quotient space $\mathbb{H}/\text{PSL}(2,\mathbb{Z})$ is shown in figure 2.2. The parameter τ is known as the modular parameter or modulus of the torus and transformations $\tau \to \tau'$ given by a $PSL(2,\mathbb{Z})$ matrix are known as modular transformations. Modular transformations are

generated by the two basic transformations

$$\tau \mapsto \tau + 1, \qquad \tau \mapsto -1/\tau \tag{2.21}$$

corresponding to the two $PSL(2,\mathbb{Z})$ matrices $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ and $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ respectively.

2.3 Functions on Riemann Surfaces

2.3.1 Holomorphic Functions.

The next thing to construct are functions on Riemann surfaces. Holomorphic and meromorphic functions play a very important role in the theory of Riemann surfaces. Below we define the notion of a function being holomorphic or meromorphic in any neighborhood of a point on a Riemann surface. We also define the notion of a harmonic function.

Let X be a Riemann surface, $p \in X$ a point in X, and f be a complex-valued function defined in a neighborhood W of p.

Definition 1 *A function $f : X \to C$ is said to be holomorphic at p if there exists a chart $\phi : U \to V$, with $p \in U$, such that the composition $f \circ \phi^{-1}$ is holomorphic at $\phi(p)$. The function f is said to be holomorphic in W if it is holomorphic at every point of W.*

If two functions f and g are both holomorphic at $p \in X$, then $f \pm g$, and fg are also holomorphic at p. If $g(p) \neq 0$, then f/g is holomorphic at p. If f is a complex-valued function on $\widehat{\mathbb{C}}$, defined in a neighborhood of ∞, then f is said to be holomorphic at ∞ if and only if $f(1/z)$ is holomorphic at $z = 0$.

2.3.2 Singularities of Functions and Meromorphic Functions.

Let X be a Riemann surface, p a point of X, and U a neighborhood of p. A punctured neighborhood of a point p is a set of the form $U - \{p\}$.

Definition 2 *Let f be a complex-valued function defined and holomorphic in a punctured neighborhood of $p \in X$. Then, the function f is said to have :*

(1) *A removable singularity at p if and only if there exists a chart $\phi : U \to V$, with $p \in U$, such that the composition $f \circ \phi^{-1}$ has a removable singularity at $\phi(p)$.*

(2) *A pole at p if and only there exists a chart $\phi : U \to V$, with $p \in U$, such that the composition $f \circ \phi^{-1}$ has a pole at $\phi(p)$.*

(3) *An essential singularity at p if and only there exists a chart $\phi : U \to V$, with $p \in U$, such that the composition $f \circ \phi^{-1}$ has an essential singularity at $\phi(p)$.*

If $\phi : U \to V$ is a chart on X with $p \in U$, considering z as the local coordinate on X near p, so that $z = \phi(x)$ for x near p, we may expand $f \circ \phi^{-1}$, which is holomorphic in a punctured neighborhood of $z_0 = \phi(p)$, in a Laurent series about z_0

$$f(\phi^{-1}(z)) = \sum_{n=-\infty}^{\infty} c_n (z - z_0)^n \ . \tag{2.22}$$

The Laurent series depends on the choice of the chart ϕ, but the nature of the singularitiy of f at a point p can be surmised from the Laurent series. The function f has a removable singularity at p if and only if its Laurent series has no terms in it with negative powers. The function f has a pole at p if and only if its Laurent series has finitely many non-zero terms with negative powers. The function f has an essential singularity at p if and only if any one of its Laurent series has infinitely many terms with negative powers. This extends the notion of a removable singularity, pole, and an essential singularity for functions of a single complex variable to functions on Riemann surfaces.

Definition 3 *A function f on X is meromorphic at a point $p \in X$ if it is either holomorphic, has a removable singularity, or has a pole, at p. The function f is said to be meromorphic on an open set W if it is meromorphic at every point of W.*

For two functions f and g holomorphic at $p \in X$, the ratio f/g is a meromorphic function at p as long as g is not identically zero in a neighborhood of p. Any function h that is meromorphic at a point $p \in X$ is locally the ratio of two holomorphic functions. If two

functions f and g are both meromorphic at $p \in X$, then $f \pm g$, and fg are also meromorphic at p. If f is a complex-valued function on $\widehat{\mathbb{C}}$, defined in a neighborhood of ∞, then f is said to be meromorphic at ∞ if and only if $f(1/z)$ is meromorphic at $z = 0$.

2.3.3 Holomorphic Maps Between Riemann Surfaces

Next we define the notion of a holomorphic mapping between Riemann surfaces. Let X and Y be two Riemann surfaces.

Definition 4 *A mapping $F : X \to Y$ is holomorphic at $p \in X$ if and only if there exist charts $\phi_1 : U_1 \to V_1$ on X with $p \in U_1$ and $\phi_2 : U_2 \to V_2$ on Y with $F(p) \in U_2$ such that the composition $\phi_2 \circ F \circ \phi_1^{-1}$ is holomorhic at $\phi_1(p)$.*

A holomorphic function then is just the special case with Y being the complex plane $\mathbb{C}$. Similar to a holomorphic function, we can also compose holomorphic maps. If $F : X \to Y$ and $G : Y \to Z$ are two holomorphic maps, then the composition $G \circ F : X \to Z$ is a holomorphic map. If $F : X \to Y$ is a holomorphic map and f is a holomorphic function on an open set $W \subset Y$, then $f \circ F$ is a holomorphic function on $F^{-1}(W)$. If $F : X \to Y$ is a holomorphic map and f is a meromorphic function on an open set $W \subset Y$, then $f \circ F$ is a meromorphic function on $F^{-1}(W)$, provided the image $F(X)$ is not a subset of the set of poles of f.

Local Normal form: Let X and Y be Riemann surfaces and $F : X \to Y$ a non-constant holomorphic map defined at a point $y \in Y$. There is a unique integer $m \geq 1$ which satisfies the property that for every chart $\phi_2 : U_2 \to V_2$ on Y centered at $F(y)$, there exists a chart $\phi_1 : U_1 \to V_1$ on X centered at y such that $\phi_2(F(\phi^{-1}(z))) = z^m$.

The multiplicity of F at a point p, denoted $\text{mult}_p(F)$, is the unique integer m such that there are local coordinates near p and $F(p)$ with F having the form $z \mapsto z^m$. The point $p \in X$ is called a ramification point for F if $\text{mult}_p(F) \geq 2$. The point $y \in Y$ is called a branch point for F if it is in the image of a ramification point for F. For each

$y \in Y$, the sum of the multiplicities of F at the points of X mapping to y is a constant, and independent of y and is called the degree of F, denoted $\deg(F)$.

2.3.4 Harmonic Functions

Finally, we define a harmonic function and mention a couple of important point about them that will be relevant to us later.

Definition 5 *A real-valued C^∞ function of two real variables, $h(x, y)$, defined on an open set $V \subset \mathbb{R}^2$ is harmonic if*

$$\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2} = 0 \tag{2.23}$$

identically on V. A complex-valued function is harmonic if and only if its real and imaginary parts are harmonic.

The Cauchy-Riemann equations for a holomorphic function immediately imply that the real and imaginary parts of any holomorphic function of $z = x + iy$ are harmonic functions of x and y. Thus, holomorphic functions are harmonic.

Harmonic functions satisfy a maximum principle.

Theorem 1 *Let f be a holomorphic function on a connected open set U of a Riemann surface X. If there is a point $p \in U$ such that $|f(x)| \leq |f(p)|$ for all $x \in U$, then f is constant on U. In particular, any holomorphic function on a compact Riemann surface X is constant.*

We will have occasion to refer to the above property in deriving an important fact about the metric on the moduli space of Seiberg-Witten theory when we study it in the next chapter.

2.4 Algebraic Curves

We understood Riemann surfaces from a topological point of view, covering them with holomorphic atlases, emphasizing their differentiable and complex structures. Another approach to Riemann surfaces is from an algebraic point of view by studying the associated algebraic curve. An algebraic curve is given by the vanishing loci of a polynomial equation. For example, we constructed two equivalent realizations of S^2 in the previous section by covering it with holomorphic atlases. We could describe S^2 just as well by its defining equation in $\mathbb{R}^3 : x^2 + y^2 + z^2 = 1,\ x, y, z \in \mathbb{R}$. Similarly, if we used complex coordinates we could describe surfaces in $\mathbb{C}^2$ by such polynomial equations. Thus, another way of realizing Riemann surfaces is as algebraic curves.

Definition 6 *An algeraic curve $\mathcal{C}$ is a subset in $\mathbb{C}^2$*

$$\mathcal{C} = \{(\mu, \lambda) \in \mathbb{C}^2 \mid \mathcal{P}(\mu, \lambda) = 0\} \tag{2.24}$$

where $\mathcal{P}(\mu, \lambda)$ is an irreducible polynomial in λ and μ

$$\mathcal{P}(\mu, \lambda) = \sum_{i=0}^{N} \sum_{j=0}^{N} p_{ij} \mu^i \lambda^j \ . \tag{2.25}$$

The curve $\mathcal{C}$ is called non-singular (or smooth) if

$$\mathrm{grad}_{\mathbb{C}}\, \mathcal{P}_{|\mathcal{P}=0} = \Big(\frac{\partial \mathcal{P}}{\partial z_1}, \frac{\partial \mathcal{P}}{\partial z_2}\Big)_{|\mathcal{P}(z_1,z_2)=0} \tag{2.26}$$

is nowhere zero on $\mathcal{C}$. The degree, d, of the curve $\mathcal{C}$ defined by $\mathcal{P}(z_1, z_2)$ is the degree of the polynomial $\mathcal{P}(z_1, z_2)$.

To make contact with the previous notion we had of Riemann surfaces, we would like to endow $\mathcal{C}$ with a complex structure. Viewing $\mathcal{C}$ as a subset of $\mathbb{C}^2$, for any non-singular curve we should be able to find a holomorphic atlas on $\mathcal{C}$ and hence endow it with a complex structure. We do that as follows:

1 Let $(z_1, z_2) \in \mathcal{C}$, be a point on the curve. If $\partial_{z_2} P(z_1, z_2) \neq 0$, by the implicit function theorem there exist open neighborhoods $\mathcal{U}_1 \in z_1$ and $\mathcal{U}_2 \in z_2$ and a holomorphic function $f_1 : \mathcal{U}_1 \to \mathcal{U}_2$ such that

$$\mathcal{C} \cap (\mathcal{U}_1 \times \mathcal{U}_2) = \{(z_1, f(z_1))\}_{z_1 \in \mathcal{U}_1}. \tag{2.27}$$

The projection $\pi_1 : (z_1, z_2) \to z_1$ maps $\mathcal{C} \cap (\mathcal{U}_1 \times \mathcal{U}_2)$ homeomorphically onto $\mathcal{U}_1$. One can take the variable z_1 as the local parameter on this chart.

2 Let $(z_1, z_2) \in \mathcal{C}$, be a point on the curve. If $\partial_{z_1} P(z_1, z_2) \neq 0$, we again have that there exist open neighborhoods $\mathcal{U}_1 \in z_1$ and $\mathcal{U}_2 \in z_2$ and a holomorphic function $f_2 : \mathcal{U}_1 \to \mathcal{U}_2$ such that

$$\mathcal{C} \cap (\mathcal{U}_1 \times \mathcal{U}_2) = \{(f(z_2), z_2)\}_{z_2 \in \mathcal{U}_2}. \tag{2.28}$$

The projection $\pi_2 : (z_1, z_2) \to z_2$ maps $\mathcal{C} \cap (\mathcal{U}_1 \times \mathcal{U}_2)$ homeomorphically onto U_2. One can take the variable z_2 as the local parameter on this chart.

Thus, we have a holomorphic atlas on $\mathcal{C}$, with the projections π_1 and π_2 giving the homeomorphisms to $\mathbb{C}$.

2.4.1 Compactification of algebraic curves

We now come to the notion of compactness and compactification. Compactness of a space is an important property. It models the properties of closed and bounded subsets of $\mathbb{R}^n$. A compact space satisfies nice properties which gives one a certain control in dealing with objects defined on it. For example, contrast the closed unit interval $[0, 1] \in \mathbb{R}$ with the open unit interval $(0, 1) \in \mathbb{R}$:

(1) All continuous functions in $[0, 1]$ are bounded while that is not true for the open interval $(0, 1)$.

(2) In $[0, 1]$ all continuous functions attain a maximum, while the same is not true for the open interval $(0, 1)$.

(3) All sequences in $[0,1]$ have convergent subsequences (Bolzano-Weistrass theorem), while that is not so for sequences in $(0,1)$.

(4) All open covers of $[0,1]$ have finite sub covers, while the open set $(0,1)$ does not possess this property.

The closed open unit interval $[0,1]$ is an example of a compact set – in fact, any closed and bounded subset of $\mathbb{R}^n$ or $\mathbb{C}^n$ is compact (Heine-Borel theorem). We can, thus, easily see how much better a compact space is to work with and the control one exercises on objects (continuous functions, etc.) defined on them. The properties of $[0,1]$ described above can be generalized to any closed and bounded subset of $\mathbb{R}^n$ or $\mathbb{C}^n$ and, in fact, taken as the definition of a compact space. Compact spaces also satisfy nice properties:

(1) If $f : \mathcal{M} \to \mathcal{N}$ is a continuous map between two topological spaces and $\mathcal{M}$ is compact, then $f(\mathcal{M})$ is compact. Thus, compactness is a topological invariant.

(2) Any closed subset of a compact space is compact.

(3) A finite union of compact spaces is compact.

(4) Product of compact sets is compact (Tychonoff's theorem).

(5) If $\mathcal{M}$ is a compact space, and $\sim$ is any equivalence relation on $\mathcal{M}$, then $\mathcal{M}/\sim$ is compact.

Since we are looking to apply our notions to algebraic curves in $\mathbb{C}^2$, we will content ourselves with this and not explore compactness of general spaces. Coming back to algebraic curves, the first fact we encounter is that algebraic curves in $\mathbb{C}^2$ are never compact. In fact, $\mathbb{C}^n$ and $\mathbb{R}^n$ are not compact. The way to remedy this situation goes by the name of compactification. One way to compactify a non-compact space is by adding the missing limit points so that every sequence now has a limit point in the compactified space. We encountered an example of this earlier when we defined the Riemann sphere : $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$. $\widehat{\mathbb{C}}$ is compact while $\mathbb{C}$ is not. We identified this with the complex projective line, $\mathbb{CP}^1$. Ideally, we would like to

find a way to repeat this process of compactification with any non-compact situations we are likely to encounter. At the very least, since we primarily wish to work with curves in $\mathbb{C}^2$, we would like to find a way to compactify $\mathbb{C}^2$ and the corresponding algebraic curves in this compactified space. That is, we wish to find the equivalent of the complex projective line $\mathbb{CP}^1$ for the case of $\mathbb{C}^2$. This is the notion of a projective space.

2.4.2 Complex projective space

Definition 7 *Complex projective space $\mathbb{CP}^n$ of dimension n is the set of all ordered $(n+1)$-tuples of complex numbers*

$$\{z = (z_0, \ldots, z_n) \in \mathbb{C}^{n+1} \mid z \neq (0, \ldots, 0)\} \tag{2.29}$$

up to the equivalence relation of scalar multiplication:

$$z \sim w \text{ if } z = \lambda w, \quad \text{for some } \lambda \in \mathbb{C}^*. \tag{2.30}$$

When $n = 1$ we get the complex projective line. For $n = 2$ we get the complex projective plane. We call coordinates $z = (z_0, \ldots, z_n) \in \mathbb{C}^{n+1}$ homogeneous coordinates on $\mathbb{CP}^n$ and denote the corresponding equivalence class by $z = [z_0, \ldots, z_n]$. As with the case of $\mathbb{CP}^1$, one can also make $\mathbb{CP}^n$ into a topological manifold by covering it with open sets:

$$\mathcal{U}_i = \{[z_0, \ldots, z_i, \ldots, z_n] \in \mathbb{CP}^n \mid z_i \neq 0\}, \quad i = 0, \ldots, n \tag{2.31}$$

and homeomorphisms to $\mathbb{C}^n$ given by

$$\phi_i : \mathcal{U}_i \to \mathbb{C}^n, \quad [z_0, \ldots, z_i, \ldots, z_n] \mapsto \Big(\frac{z_0}{z_i}, \ldots \frac{z_{i-1}}{z_i}, \frac{z_{i+1}}{z_i}, \ldots, \frac{z_n}{z_i}\Big). \tag{2.32}$$

The complement of $U_i \in \mathbb{CP}^n$ is the set

$$\{[z_0, \ldots, z_i, \ldots, z_n] \in \mathbb{CP}^n \mid z_i = 0\} \tag{2.33}$$

known as a hyperplane. This set can be identified with $\mathbb{CP}^{n-1}$ in the obvious way. Continuing this process now in $\mathbb{CP}^{n-1}$, we can inductively identify the hyperplanes with lower

dimensional complex projective spaces. Reversing this process we can build $\mathbb{CP}^n$ starting with $\mathbb{CP}^0$, which is just a point. Taking this point (as the point at infinity) together with a copy of $\mathbb{C}$ gives us $\mathbb{CP}^1$. Now, $\mathbb{CP}^2$ is obtained by taking $\mathbb{CP}^1$ as the "line at infinity" together with a copy of $\mathbb{C}^2$. Continuing inductively, we see that we construct $\mathbb{CP}^n$ as a copy of $\mathbb{C}^n$ together with a copy of $\mathbb{CP}^{n-1}$ at infinity. It remains to see that this construction gives us a compact space, i.e. $\mathbb{CP}^n$ is compact.

Proposition 1 *The space $\mathbb{CP}^n$ is compact.*

Our strategy will be to find a suitable compact subset of $\mathbb{C}^{n+1}$, and a continuous map from this subset to $\mathbb{CP}^n$. A closed and bounded subset of $\mathbb{C}^{n+1}$, is compact by the Heine-Borel theorem, which states that a subset of $\mathbb{R}^n$ or $\mathbb{C}^n$ that is closed and bounded is compact, and the image of a continious map from a compact set is again compact. This will lead to the desired result. Consider the $(2n+1)$ sphere

$$S^{2n+1} = \{(z_0, \ldots, z_n) \in \mathbb{C}^{n+1} : |z_0|^2 + \ldots |z_n|^2 = 1\}$$

which is a closed and bounded subset of $\mathbb{C}^{n+1}$ and hence compact . The map

$$\Pi : \mathbb{C}^{n+1} - \{0\} \to \mathbb{CP}^n : (z_0, \ldots, z_n) \mapsto [z_0, \ldots, z_n] \tag{2.34}$$

restricted to $S^{2n+1} \in \mathbb{C}^{n+1}$ is continuous and hence the image is again compact.
Now, for $[z_0, \ldots, z_n] \in \mathbb{CP}^n$,

$$\lambda = |z_0|^2 + \ldots |z_n|^2 > 0$$

and so

$$[z_0, \ldots, z_n] = [\lambda^{-\frac{1}{2}} z_0, \ldots, \lambda^{-\frac{1}{2}} z_n]$$

But,

$$|\lambda^{-\frac{1}{2}} z_0|^2 + \ldots |\lambda^{-\frac{1}{2}} z_n|^2 = 1$$

so $[z_0, \ldots, z_n] \in \Pi(S^{2n+1})$. Thus, $\Pi : S^{2n+1} \to \mathbb{CP}^n$ is surjective and hence we have that $\mathbb{CP}^n$ is compact.

Complex projective curves in $\mathbb{CP}^2$

Having constructed our compactification of $\mathbb{C}^2$, we would like to study our previously defined notions of algebraic curves living in this new setting. Recall that an algebraic curve in $\mathbb{C}^2$ was defined as the set of zeros of a non-constant irreducible polynomial of z_1 and z_2 with no repeated factors. So, our first order of business is to see what polynomials have a well defined zero locus in $\mathbb{CP}^n$. Then, we can look at their vanishing loci and get projective curves out of them. Now, coordinates in $\mathbb{CP}^n$ are equivalence classes of $(n+1)$-tuples (not all zero) that are equivalent up to multiplication by a non-zero constant $\lambda \in \mathbb{C}^*$. Thus, to have a consistent vanishing locus, a polynomial that vanishes at a point $z = [z_0, \ldots, z_n] \in \mathbb{CP}^n$, must also vanish for all points in the equivalence class that the point z is in, i.e.

$$P([z_0, \ldots, z_n]) = 0 \Leftrightarrow P([\lambda z_0, \ldots, \lambda z_n]) = 0, \quad \forall \lambda \in \mathbb{C}^*.$$

This is achieved by defining the concept of homogeneous polynomials.

Definition 8 *A polynomial $P(z_0, \ldots, z_n)$ in the variables $(z_0, \ldots, z_n)$ is called homogeneous of degree d if*

$$P(\lambda z_0, \ldots, \lambda z_n) = \lambda^d P(z_0, \ldots, z_n) \tag{2.35}$$

for all $\lambda \in \mathbb{C}$.

Corresponding to the homogeneous polynomial $P(z)$ of degree d on the open set $\mathcal{U}_i$ we have a polynomial p in $\mathbb{C}[z_1, \ldots, z_n]$ given by

$$P([z_0, \ldots, z_n]) = z_i^d p\left(\frac{z_0}{z_i}, \ldots, \frac{z_{i-1}}{z_i}, \frac{z_{i+1}}{z_i}, \ldots, \frac{z_n}{z_i}\right) .$$

Armed with the appropriate set of polynomials we are ready to define a projective curve in $\mathbb{CP}^2$ along the lines of an algebraic curve in $\mathbb{C}^2$.

Definition 9 *Let $P(z_0, z_1, z_2)$ be a nonconstant homogeneous polynomial in the variables (z_0, z_1, z_2) with complex coefficients and with no repeated factors. The projective curve $\hat{\mathcal{C}}$ definied by $P(z_0, z_1, z_2)$ in $\mathbb{CP}^2$ is given by*

$$\hat{\mathcal{C}} = \{[z_0, z_1, z_2] \in \mathbb{CP}^2 : P(z_0, z_1, z_2) = 0\}. \tag{2.36}$$

Other notions about complex algebraic curves generalize to complex projective curves as well. A point $[z_0, z_1, z_2]$ on a projective curve $\hat{\mathcal{C}} \in \mathbb{CP}^2$ defined by the homogeneous polynomial $P(z_0, z_1, z_2)$ is called singular if

$$\frac{\partial P}{\partial z_0}(z_0, z_1, z_2) = \frac{\partial P}{\partial z_1}(z_0, z_1, z_2) = \frac{\partial P}{\partial z_2}(z_0, z_1, z_2) = 0. \tag{2.37}$$

If the curve $\hat{\mathcal{C}}$ has no singular points, it is called nonsingular. The degree of a projective curve $\hat{\mathcal{C}}$ in $\mathbb{CP}^2$ defined by a homogeneous polynomial $P(z_0, z_1, z_2)$ is the same as the degree of the polynomial $P(z_0, z_1, z_2)$. The projective curve $\hat{\mathcal{C}}$ is called irreducible if the polynomial $P(z_0, z_1, z_2)$ defining it is irreducible.

Let us briefly look at the relationship between algebraic curves defined in section 2.4 and projective curves. One can obtain a projective curve $\hat{\mathcal{C}}$ from an algebraic curve $\mathcal{C}$ by adding "points at infinity", which is the process of compactification. Using (2.32) we can identify $\mathbb{C}^2$ with the open set

$$U = \{[z_0, z_1, z_2] \in \mathbb{CP}^2 : z_2 \neq 0\} \tag{2.38}$$

in $\mathbb{CP}^2$ where the homeomorphism $\phi : U \to \mathbb{C}^2$ is given by

$$\phi\left([z_0, z_1, z_2]\right) = \left(\frac{z_0}{z_2} = \mu, \frac{z_1}{z_2} = \lambda\right) \tag{2.39}$$

and its inverse,

$$\phi^{-1}\left(\mu, \lambda\right) = [\mu, \lambda, 1]. \tag{2.40}$$

Further when $z_2 = 0$, we get the complement of this open set in $\mathbb{CP}^2$:

$$U^c = \{[z_0, z_1, z_2] \in \mathbb{CP}^2 : z_2 = 0\}. \tag{2.41}$$

The set U^c can be identified with $\mathbb{CP}^1$ through the map $[z_0, z_1, 0] \mapsto [z_0, z_1]$. This copy of $\mathbb{CP}^1$ is thought of as the line "at infinity" in $\mathbb{CP}^2$. Thus we can see that the projective plane $\mathbb{CP}^2$ is the disjoint union of a copy of $\mathbb{C}^2$ and the projective line $\mathbb{CP}^1$. Consider now the projective curve $\hat{\mathcal{C}}$ defined by a non-constant homogenous polynomial $P\left([z_0, z_1, z_2]\right)$ of degree d. Identifying U with $\mathbb{C}^2$ as above, the intersection of $\hat{\mathcal{C}}$ and U gives the algebraic

curve in $\mathbb{C}^2$ defined by the inhomogenous polynomial $P([z_0, z_1, 1])$ in two variables. If $\hat{\mathcal{C}}$ does not contain the line $z_2 = 0$, the polynomial $P([z_0, z_1, 1])$ is of degree d.
In the other direction, an algebraic curve $\mathcal{C}$ defined by an inhomogenous polynomial of degree d in two vairables $\mathcal{P}(z_0, z_1)$, can be considered as the intersection of U with the projective curve $\hat{\mathcal{C}}$ defined by the homogeneous polynomial

$$P([z_0, z_1, z_2]) = z_2^d \, \mathcal{P}\left(\frac{z_0}{z_2}, \frac{z_1}{z_2}\right) . \tag{2.42}$$

The intersection of this projective curve with the line at infinity, $z_2 = 0$, is given by the vanishing of the homogeneous polynomial

$$P([z_0, z_1, 0]) = 0, \quad [z_0, z_1, 0] \in \mathbb{CP}^2 \tag{2.43}$$

in z_0 and z_1. We now show that this condition determines a set of points.

A nonzero homogeneous polynomial of degree d in two variables with complex coefficients $c_0, \ldots, c_d \in \mathbb{C}$,

$$P([z_0, z_1]) = \sum_{r=0}^{d} c_r z_0^r \, z_1^{d-r}, \tag{2.44}$$

can always be factored into a product of linear factors,

$$P([z_0, z_1]) = \prod_{i=1}^{d} (\alpha_i z_0 + \beta_i z_1) , \tag{2.45}$$

for some $\alpha_i, \beta_i \in \mathbb{C}$. This can be seen as follows. Let m be the largest element of $\{0, \ldots, d\}$ such that the coefficient c_m is nonzero. Writing $P([z_0, z_1])$ as

$$P([z_0, z_1]) = \sum_{r=0}^{d} c_r z_0^r \, z_1^{d-r} = z_1^d \sum_{r=0}^{d} c_r \left(\frac{z_0}{z_1}\right)^r , \tag{2.46}$$

we can factorize $\sum_{r=0}^{d} c_r \, (z_0/z_1)^r$ in the variable z_0/z_1 as a polynomial of degree m with complex coefficients

$$\sum_{r=0}^{d} c_r \left(\frac{z_0}{z_1}\right)^r = c_m \prod_{i=1}^{m} \left(\frac{z_0}{z_1} - \gamma_i\right) , \tag{2.47}$$

for some $\gamma_1, \ldots, \gamma_m \in \mathbb{C}$. Then,

$$\begin{aligned} P([z_0, z_1]) &= c_m z_1^d \prod_{i=1}^{m} \left(\frac{z_0}{z_1} - \gamma_i \right) \\ &= c_m z_1^{d-e} \prod_{i=1}^{m} (z_0 - \gamma_i z_1), \end{aligned} \tag{2.48}$$

and we have the desired form. In $\mathbb{CP}^2$, the factors $\alpha_i z_0 + \beta_i z_1 = 0$ correspond to the points $[-\beta_i, \alpha_i, 0]$, and thus the intersection of the projective curve $\hat{\mathcal{C}}$ with the line at infinity is a set of points. These are the points of $\hat{\mathcal{C}} - \mathcal{C}$. We will see below that the projective curve $\hat{\mathcal{C}}$ is compact. Thus $\hat{\mathcal{C}}$ is the desired compactifiation of $\mathcal{C}$. To complete our discussion, we would like to see the compactness of projective curves.

Proposition 2 *A projective curve*

$$\hat{\mathcal{C}} = \{[z_0, z_1, z_2] \in \mathbb{CP}^2 : P(z_0, z_1, z_2) = 0\} \tag{2.49}$$

in $\mathbb{CP}^2$ *is compact.*

We show that $\hat{\mathcal{C}}$ is a closed subset of $\mathbb{CP}^2$. Since any closed subset of a compact set is again compact, $\hat{\mathcal{C}}$ is compact. Considering the map (2.34), we can give $\mathbb{CP}^n$ the quotient topology induced from the usual topology on $\mathbb{C}^{n+1} - \{0\}$. Then, a subset $A \subset \mathbb{CP}^n$ is open if and only if $\Pi^{-1}(A)$ is an open subset of $\mathbb{C}^{n+1} - \{0\}$, and a subset $B \subset \mathbb{C}P^n$ is closed if and only if $\Pi^{-1}(B)$ is a closed subset of $\mathbb{C}^{n+1} - \{0\}$.

Thus, to show that $\hat{\mathcal{C}}$ is a closed subset of $\mathbb{CP}^2$, we require that

$$\Pi^{-1}(\hat{\mathcal{C}}) = \{(z_0, z_1, z_2) \in \mathbb{C}^3 - \{0\} : P(z_0, z_1, z_2) = 0\} \tag{2.50}$$

be a closed subset of $\mathbb{C}^3 - \{0\}$. Since polynomials are continuous, this is obviously true and hence the curve $\hat{\mathcal{C}}$ is compact.

2.5 Covering spaces and Monodromy

In this section we understand the concept of monodromy matrices associated to Riemann surfaces. The idea of monodromy arises in dealing with multivaluedness. We will start by

understanding the concept of the fundamental group.

2.5.1 Fundamental Group

Let us first fix a complex manifold $\mathcal{M}$. The fundamental group of the manifold $\mathcal{M}$ is defined with respect to a given base point $p \in \mathcal{M}$. Let us fix a base point $p \in \mathcal{M}$. A *path* on $\mathcal{M}$ is a continuous map $\gamma : [0,1] \to \mathcal{M}$ with initial point $\gamma(0) = x_0$ and final point $\gamma(1) = x_1$. If $\gamma_1(t)$ and $\gamma_2(t)$ are two paths such that $\gamma_1(1) = \gamma_2(0)$, then we can define a product on the space of such paths as

$$\gamma_1 * \gamma_2(t) = \begin{cases} \gamma_1(2t) & 0 \leq t \leq \frac{1}{2} \\ \gamma_2(2t-1) & \frac{1}{2} \leq t \leq 1 \end{cases} . \tag{2.51}$$

This is equivalent to traversing along $\gamma_1(I)$ in the first half followed by $\gamma_2(I)$ in the other half, where $I = [0,1]$ is the unit interval. A *loop* based at a point x is a path such that $\gamma(0) = \gamma(1) = x$. On a smooth manifold, some loops can be deformed into each other continuously and this leads to the notion of homotopy of loops. Two given loops γ_1 and γ_2 based at the same point p are said to be *homotopic* to each other if there is a continuous map $F : [0,1] \times [0,1] \to \mathcal{M}$ such that $F(0,t) = \gamma_1(t)$ and $F(1,t) = \gamma_2(t)$ for all $t \in I$ and $F(s,0) = F(s,1) = p$ for all $s \in I$. The connecting map F is called the homotopy between γ_1 and γ_2. Obviously, such a map is possible only when there is no obstruction (in the form of holes in the space) to continuously deforming one loop to the other. Homotopy is an equivalence relation on the set of loops based at a given point. If there is an obstruction to deforming one loop into another, they would belong to different equivalence classes and, every obstruction would give rise to new equivalence classes of loops. One can use (2.51) to give a product structure on the space of equivalence classes of loops on $\mathcal{M}$ making this space into a group. The unit element for this group is $[e_p]$, the class of the constant loop at p defined as $e_p : I \to \mathcal{M}$, with $e_p(s) = p$ for $s \in I$. The group of equivalence classes of loops based at $p \in \mathcal{M}$ is known as the fundamental group denoted $\pi_1(\mathcal{M}, p)$. Intuitively, the fundamental group gives one information about the holes in the space. Nevertheless, the definition of the fundamental group would seem to imply it depends on the choice of the

base point $p \in \mathcal{M}$. For the fundamental group to be a topological invariant, however, we should require that the fundamental groups based at two different points on the manifold $\mathcal{M}$ be the same, i.e. $\pi_1(\mathcal{M}, p) \equiv \pi_1(\mathcal{M}, q)$ for any $p, q \in \mathcal{M}$. To see that this is indeed so, let γ be the path from p to q. Then, for each class $[\alpha] \in \pi_1(\mathcal{M}, p)$, we get the isomorphism between $\pi_1(\mathcal{M}, p)$ and $\pi_1(\mathcal{M}, q)$ given by $[\alpha] \mapsto [\gamma^{-1}][\alpha][\gamma] \in \pi_1(\mathcal{M}, q)$. Thus, for a path connected manifold, the fundamental group is a topological invariant – i.e. it depends only on the topology of the manifold in question. A path connected space such that any loop in the space can be contracted to a point is called a simply connected space. For example, $\mathbb{R}^n$ is a simply connected space. Let us see some examples of fundamental groups of some well known spaces.

2.5.2 Examples of fundamental groups

$\mathbb{R}^n$

To find the fundamental group of $\mathbb{R}^n$, we observe that $\mathbb{R}^n$ is path-connected, and any loop in $\mathbb{R}^n$ can be contracted to a point. A loop based at a point, say $x_0 \in \mathbb{R}^n$ is homotopy equivalent to any other loop based at x_0. This is because there are no obstructions (a hole, for example) in the space to finding a homotopy map between any two loops based at x_0. Thus, all loops based at the point x_0 are homotopy equivalent to each other, and there is just one equivalence class of loops and hence just one element in the fundamental group. The fundamental group of $\mathbb{R}^n$ is $\pi_1(\mathbb{R}^n, x_0) = \{e\}$.

We can see that the fundamental group of any simply connected space is the same as the fundamental group of a point, and hence trivial.

Circle – S^1

The circle has a hole in it, so we should guess that its fundamental group would not be trivial like in the case of $\mathbb{R}^n$. A loop in S^1 that goes around it cannot be equivalent to the constant loop like in the case of $\mathbb{R}^n$ because of the presence of the hole. In fact, any loop

that goes around the circle n number of times cannot be homotopy equivalent to a loop that goes around the circle m number of times unless $n = m$ by the same argument. This suggests that the integer n characterizes the equivalence classes of homotopic loops on the circle. The fundamental group of the circle is $\pi_1(S^1) \cong \mathbb{Z}$.

Torus – T^2

Let $T^2 = S^1 \times S^1$ be a torus. Intuitively it would require us two independent integers to characterize the loops going around each of the S^1s. This suggests the fundamental group of the torus to be

$$\pi_1(T^2) \cong \pi_1(S^1) \times \pi_1(S^1) \cong \mathbb{Z}^2. \tag{2.52}$$

Extending the logic to the case of the n-dimensional torus, T^n, given by

$$T^n = \underbrace{S^1 \times S^1 \times \cdots \times S^1}_{n \text{ factors}} \tag{2.53}$$

we have the fundamental group of T^n to be

$$\pi_1(T^n) \cong \mathbb{Z}^n. \tag{2.54}$$

In fact, for two arcwise connected topological spaces X and Y, $\pi_1(X \times Y, (x_0, y_0)) \cong \pi_1(X, x_0) \times \pi_1(Y, y_0)$. Hence, for example, the fundamental group of the space $X = S^1 \times \mathbb{R}$ (a cylinder) is given by

$$\pi_1(X) \cong \mathbb{Z} \oplus \{e\} \cong \mathbb{Z}. \tag{2.55}$$

We can also see that the space $X = \mathbb{R}^2 \setminus \{(0, 0)\}$ which is the real plane with the origin removed has a hole (where the point $\{(0, 0)\}$ used to be in $\mathbb{R}^2$), and hence the fundamental group of X should be isomorphic to that of the circle S^1.

2.5.3 Branched coverings

Let Y be a connected real manifold. A covering space of Y is a continuous map $F : X \to Y$ such that F is onto, and for each point $y \in Y$ there is a neighborhood W of y in Y

such that $F^{-1}(W)$ consists of a disjoint union of open sets X_α, each mapping onto W homeomorphically by F. The number of preimages of a point in Y is called the degree of the covering. Two such covers, $F_1 : X_1 \to Y$ and $F_2 : X_2 \to Y$ are said to be isomorphic if there is a homeomorphism $G : X_1 \to X_2$ such that $F_2 \circ G = F_1$. We can consider the case when X and Y are Riemann surfaces and the mapping F a non-constant holomorphic map from X to Y. When we consider Y to be $\mathbb{CP}^1$, we get branched coverings of the Riemann sphere.

A branched covering of the Riemann sphere is a pair $(\mathcal{C}, f)$ where $\mathcal{C}$ is a compact connected Riemann surface and $f : \mathcal{C} \to \mathbb{CP}^1$ is a non-constant holomorphic map. Two branched coverings $(\mathcal{C}, f)$ and $(\mathcal{C}', f')$ are equivalent if there is a biholomorphism[1] $g : \mathcal{C} \to \mathcal{C}'$ such that $f = f' \circ g$. Every compact Riemann surface can arise as a branched covering of $\mathbb{CP}^1$. Ramification points and branch points of the covering are the ramification and branch points of the function f (see Section 2.3.3). The degree of the function f is called the degree of the branched covering.

For a branched covering $(\mathcal{C}, f)$, and any point $q \in \mathbb{CP}^1$ that is not a branch point of f, there exists a neighbourhood $U \subset \mathbb{CP}^1$ of q and each connected component of $f^{-1}(U)$ is homeomorphic to U by f.

Outside the branch points the branched covering is a (topological) covering, i.e. if $A = \{\text{set of branch points}\}$ then the restricted map

$$f_| : \mathcal{C} \setminus f^{-1}(A) \to \mathbb{CP}^1 \setminus A \tag{2.56}$$

is such that for any $q \in \mathbb{CP}^1 \setminus A$ there exists a neighborhood U of q such that $f^{-1}(U) \subset \mathcal{C} \setminus f^{-1}(A)$ is homeomorphic to $U \times S$ for a discrete set S. The connected components of the preimage $f^{-1}(U)$ are called the sheets of the covering over U.

For example, we can consider the covering of $\mathbb{CP}^1$ by the nonsingular projective curve $\mathcal{C} \subset \mathbb{CP}^2$ defined by the homogeneous polynomial $P(x, y, z)$ of degree $d > 1$. We may

[1] A holomorphic mapping f is said to be biholomorphic if it is one-to-one and the inverse f^{-1} is holomorphic as well.

assume that the point $[0,1,0] \notin \mathcal{C}$. Then, we can define the map $f : \mathcal{C} \to \mathbb{CP}^1$ given by

$$f[x,y,z] = [x,z]. \tag{2.57}$$

For example, if we consider the equation $y^2 = xz$, the preimage of $[x,z]$ generally consists of 2 points unless $x = 0$ or $z = 0$. When $x = 0$ or $z = 0$, the corresponding points $[0,1]$ and $[1,0]$ in $\mathbb{CP}^1$ have just one point in $\mathcal{C}$ corresponding to them. The covering ϕ is branched over the points $[0,1]$ and $[1,0]$ in $\mathbb{CP}^1$.

More generally, for the covering by the projective curve $\mathcal{C}$ given by the degree d polynomial $P(x,y,z)$, a point $[a,b,c] \in \mathcal{C}$ is a ramification point of ϕ if the order $\nu_\phi[a,b,c]$ of the zero of the polynomial $P(a,y,c)$ in y at $y = b$ is greater than 1. Any nonsingular projective curve $\mathcal{C} \subset \mathbb{CP}^2$ of degree $d > 1$ can be viewed as a branched covering of $\mathbb{CP}^1$.

2.5.4 Monodromy

In this section we study the idea of monodromy associated to a finite covering. Consider the unit circle, S^1. Let us take a string and wind it around S^1 counterclockwise, starting at a point, say, $z_0 \in S^1$. When the string goes around S^1 once and comes back to z_0 the length of string we have used up is 2π. Thus, although we are back at the point z_0 on S^1, the point on the string that corresponds to it is 2π and not 0. Going around n times and coming back to the point z_0, we use $2\pi n$ length of string. The number n is called the winding number. A winding in the clockwise direction is taken with the negative sign. Thus, we see that going around S^1 and coming back to z_0 does not bring us back to 0 on the string. Instead going around z_0 adds 2π to the length of wound string. We see how a quantity (the length of the string) can have multiple values at the same point $z_0 \in S^1$ in going around a loop on this manifold (S^1). We also understand clearly how this is a result of the topology of the manifold – specifically the fact that S^1 contains a hole and any loop around this can not be shrunk back to a point. This can of course be made more precise by employing the idea of the fundamental group. The winding number, which tells us how much string has been used, is then given as the map from the fundamental group to the

integers: $n : \pi_1(S^1, z_0) \to \mathbb{Z}$. We study the monodromy of a finite covering in this spirit.

Monodromy of a Finite Covering

Let $F : X \to Y$ be a covering of a connected real manifold Y of finite degree d. Over each $y \in Y$, the fiber $F^{-1}(y)$ consists of d points, $\{x_1, \ldots, x_d\}$. Every loop γ in Y based at y can be lifted to d paths $\tilde{\gamma_1}, \ldots, \tilde{\gamma_d}$, where the lift $\tilde{\gamma_i}$ is the unique lift of γ that starts at x_i, i.e. $\tilde{\gamma_i}(0) = x_i$, for every i. The end points, $\tilde{\gamma_i}(1)$ for each of the loops also lie in the preimage set $F^{-1}(y)$ over y and hence each must be one of the x_j for some j. We denote the end points of the loops, $\tilde{\gamma_i}$ by $x_{\sigma(i)}$, where σ is a permutation of the indices $\{1, \ldots, d\}$. The permutation σ depends only on the homotopy class of the loop γ, and therefore we have a group homomorphism

$$\rho : \pi_1(Y, y) \to S_d \tag{2.58}$$

where S_d is the symmetric group of all permutations on d elements. The group homomorphism ρ of (2.58) is called the monodromy representation of the degree d covering map $F : X \to Y$. If X is connected, then the image of ρ is a transitive subgroup of S_d.

The Monodromy of a Holomorphic Map

We can also apply the above idea of monodromy representations to the case of a holomorphic nonconstant map between compact Riemann surfaces. In this case, however, due to the presence of branch points and ramification, the map is not in general a covering map. Consider the nonconstant holomorphic map $F : X \to Y$, where X and Y are compact Riemann surfaces. Let $R \subset X$ and $B = F(R) \subset Y$, be the finite set of ramification points, and branch points of F respectively. Let $V = Y - B$ and let $U = X - F^{-1}(B)$. From Y, we remove all the branch points, and from X, we remove all the ramification points, as well as any point that is mapped to a branch point under F. That is, we remove all the points, all of which are not necessarily ramification points, in the same fiber of F as a ramification point.

The restriction of the map F to U and V, $F\mid_U: U \to V$, is a true covering map of degree d, since now, for any $v \in V$, the preimage set $F^{-1}(v)$ consists of d distinct points, each with multiplicity one for the holomorphic map F. This covering map leads to a monodromy representation $\rho : \pi_1(V, q) \to S_d$, called the monodromy represenation of the holomorphic map F. Further, since X is connected, so is the open set U and the image is a transitive subgroup of S_d.

In the context of our study of branched coverings of $\mathbb{CP}^1$ and the fundamental group, multivaluedness and the monodromy group arise in the following way. Consider the branched covering $f : \mathcal{C} \to \mathbb{CP}^1$ of $\mathbb{CP}^1$ branched at the points $\{a_1, \ldots, a_n\} \subset \mathbb{CP}^1$. For any loop γ in $\mathbb{CP}^1$ based at the point $q \in \mathbb{CP}^1 \setminus \{a_1, \ldots, a_n\}$, there is a unique lift $\widetilde{\gamma}_p$ of γ to a path in $\mathcal{C}$ which starts at $p \in \phi^{-1}(q)$. If we consider a loop $\gamma_{i,q}$ based at $q \in \mathbb{CP}^1$ and going once around the branch point a_i, the lift of $\gamma_{i,q}$ in $\mathcal{C}$, $\widetilde{\gamma}_{i,p}$, ends in $\phi^{-1}(q)$ but not necessarily at the point p. Let the end point of the loop be at $M_{\gamma_{i,q}}(p)$. Thus, going around a loop might take us to a different sheet in the covering when one goes around a loop enclosing a branch point. This is the source of our multivaluedness. The function $M_{\gamma_{i,q}} : \phi^{-1}(q) \to \phi^{-1}(q)$ encodes the change in going around the loop $\gamma_{i,q} \in \mathbb{CP}^1$. This map is also invertible since the loop $\gamma_{i,q}^{-1}$ is the loop $\gamma_{i,q}$ with its direction reversed. This map is also invariant under homotopy and compatible with composition of loops based at the point q. Thus, we have a representation of the fundamental group $\pi_1(\mathbb{CP}^1, q)$, which is the monodromy representation for the covering $\phi : \mathcal{C} \to \mathbb{CP}^1$ with the given branch points data.

2.6 First Homology Group

To define the first homology group of a compact Riemann surface X of genus g, we form the abelian groups out of formal sums of points, oriented curves, and oriented domains of X, with integer coefficients. The formal sums of points, $\sum n_i P_i$, oriented curves, γ_i

$$\gamma = \sum n_i \gamma_i, \tag{2.59}$$

and oriented domains

$$D = \sum n_i D_i \tag{2.60}$$

with integer coefficients $n_i \in \mathbb{Z}$ form the abelian groups C_0, C_1 and C_2 respectively. The elements of these groups are called 0-chains, 1-chains, and 2-chains respectively. The boundary operator ∂ that maps the elements to their oriented boundaries defines the group homomorphisms

$$\partial_r : C_r \to C_{r-1} \ . \tag{2.61}$$

Let us look at the group C_1. An oriented loop γ (i.e. $\partial_1 \gamma = 0$) is called a *cycle*, and a $\gamma \in C_1$ of the form $\gamma = \partial_2 D$ is called a *boundary*. The subgroups of cycles and boundaries are denoted

$$Z = \{\gamma \in C_1 \mid \partial_1 \gamma = 0\}, \qquad B = \{\gamma \in C_1 \mid \gamma = \partial_2 \delta, \text{ for some } \delta \in C_2\}. \tag{2.62}$$

The subgroup Z is given by the kernel of $\partial_1 : C_1 \to C_0$ and the subgroup B is the image of $\partial_2 : C_2 \to C_1$. Every boundary is a cycle and the set of boundaries is a subset of the set of cycles, i.e. $B \subset Z \subset C_1$. Two elements of C_1 that differ by a boundary are called homologous:

$$\gamma_1, \gamma_2 \in C_1, \quad \gamma_1 \sim \gamma_2 \iff \gamma_1 - \gamma_2 \in B \ . \tag{2.63}$$

The elements of Z that are not boundaries play an important role. The first homology group of X is defined as the factor group

$$H_1(X, \mathbb{Z}) = Z/B \ . \tag{2.64}$$

Every element of $H_1(X, \mathbb{Z})$ can be represented by a smooth cycle without self-intersection. Moreover, given two elements of $H_1(X, \mathbb{Z})$, one can represent them by smooth cycles that intersect transversally in a finite number of points. Let γ_1 and γ_2 be two curves that intersect transversally at a point p. To this point one associates the number $(\gamma_1 \circ \gamma_2)_p = \pm 1$, where the sign is determined by the orientation of γ_1 and γ_2 as shown in the figure 2.3. For two smooth cycles γ_1, γ_2 intersecting transversally in finitely many points, the intersection

number of γ_1 and γ_2 is defined by

$$\gamma_1 \circ \gamma_2 = \sum_{p \in \gamma_1 \cap \gamma_2} (\gamma_1 \circ \gamma_2)_p \ . \tag{2.65}$$

To define the intersection number for homology classes $\gamma, \gamma' \in H_1(X, \mathbb{Z})$, represent them as

$$\gamma = \sum_i n_i \gamma_i, \qquad \gamma' = \sum_j m_j \gamma'_j, \tag{2.66}$$

where γ_i, γ'_j are smooth curves intersecting transversally. Define

$$\gamma \circ \gamma' = \sum_{ij} n_i m_j \gamma_i \circ \gamma'_j \ . \tag{2.67}$$

From this, we have a bilinear skew-symmetric map

$$\circ : H_1(X, \mathbb{Z}) \times H_1(X, \mathbb{Z}) \to \mathbb{Z} \tag{2.68}$$

called the intersection number on the elements of $H_1(X, \mathbb{Z})$. The first homology group $H_1(X, \mathbb{Z})$ of a genus $g \geq 1$ Riemann surface is isomorphic to the additive group $\mathbb{Z}^{2g}$ and is generated by cycles, denoted $a_1, b_1, \ldots, a_g, b_g$. The intersection numbers of these cycles are given by

$$a_i \circ b_i = \delta_{ij}, \qquad a_i \circ a_i = b_i \circ b_i = 0 \ . \tag{2.69}$$

A basis of g cycles satisfying (2.69) is called canonical basis of cycles. A canonical basis of cycles is unique up to the action of a $\mathrm{Sp}(g, \mathbb{Z})$ matrix. We can see this by representing the canonical basis as a $2g$ dimensional vector

$$\begin{pmatrix} a \\ b \end{pmatrix}, \qquad a = \begin{pmatrix} a_1 \\ \vdots \\ a_g \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_g \end{pmatrix}, \tag{2.70}$$

and writing the intersections numbers as

$$\begin{pmatrix} a \\ b \end{pmatrix} \circ \begin{pmatrix} a & b \end{pmatrix} = J, \qquad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \tag{2.71}$$

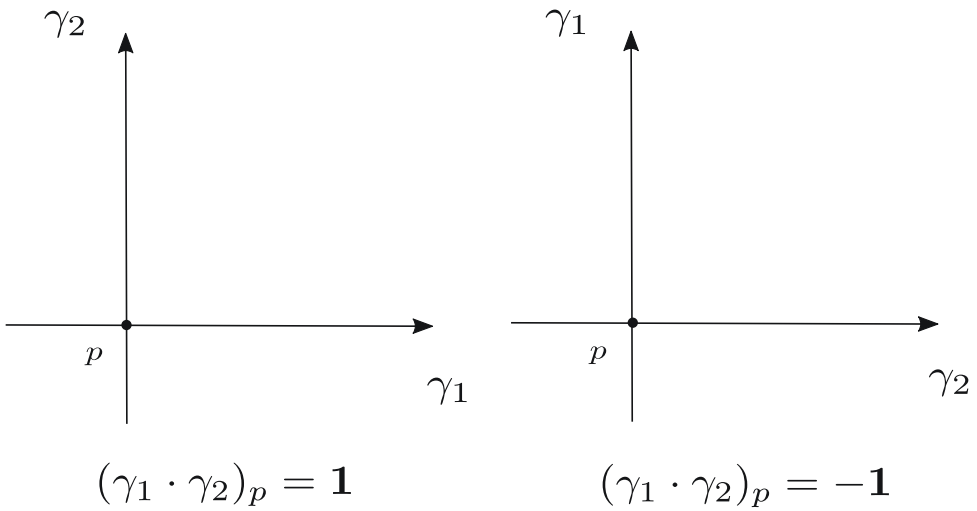

Figure 2.3: Intersection number at a point

where I is the g-dimensional identity matrix. For another canonical basis $(\tilde{a}, \tilde{b})$ related to (a, b) by

$$\begin{pmatrix} \tilde{a} \\ \tilde{b} \end{pmatrix} = M \begin{pmatrix} a \\ b \end{pmatrix}, \quad M \in \mathrm{GL}(2g, \mathbb{Z}), \tag{2.72}$$

we have

$$J = \begin{pmatrix} \tilde{a} \\ \tilde{b} \end{pmatrix} \circ \begin{pmatrix} \tilde{a} & \tilde{b} \end{pmatrix} = M \begin{pmatrix} a \\ b \end{pmatrix} \circ \begin{pmatrix} a & b \end{pmatrix} M^T = M J M^T \ . \tag{2.73}$$

Thus, the basis $(\tilde{a}, \tilde{b})$ is canonical if and only if M is a symplectic matrix, i.e. $M \in \mathrm{Sp}(g, \mathbb{Z})$.

2.7 Differential Forms and Integration

We finally introduce differential forms and integration on Riemann surfaces. This helps us introduce the important tool of contour integration on Riemann surfaces. Let X be a Riemann surface of genus g, and $z = x + iy$ be a local parameter in some domain U. Any differential 1-form on X can locally be written in the form $\omega = u(x, y)\mathrm{d}x + iv(x, y)\mathrm{d}y$. In terms of the basis $\mathrm{d}z = \mathrm{d}x + i\mathrm{d}y$ and $\mathrm{d}\bar{z} = \mathrm{d}x - i\mathrm{d}y$, we can rewrite the differential 1-form as $\omega(z, \bar{z}) = f(z, \bar{z})\mathrm{d}z + g(z, \bar{z})\mathrm{d}\bar{z}$. The parts $f\mathrm{d}z$ and $g\mathrm{d}\bar{z}$ are called the $(1, 0)$- and the $(0, 1)$-forms respectively.

Definition 10 *A differential form ω on X is called a holomorphic differential (or a differential of the first kind) if it can be expressed locally in the form*

$$\omega = f(z)\mathrm{d}z \tag{2.74}$$

where f is a holomorphic function of the local parameter z.

The space of all holomorphic differentials on X constitutes a linear vector space. The dimension of the space of holomorphic differentials on a compact Riemann surface is equal to its genus (for a proof, see [9]). For example, on a hyperelliptic Riemann surface

$$\mu^2 = \prod_{i=1}^{N} (\lambda - \lambda_i) \qquad N \geq 3, \quad \lambda_i \neq \lambda_j, \tag{2.75}$$

with $N = 2g + 2$ or $N = 2g + 1$, the basis of holomorphic differentials is given by

$$\omega_j = \frac{\lambda^{j-1}}{\mu}\mathrm{d}\lambda, \qquad j = 1, \ldots, g \ . \tag{2.76}$$

When $N = 3$ or 4 the curve (2.75) is called an elliptic curve. For $\omega = f\mathrm{d}z + g\mathrm{d}\bar{z}$, we have,

$$\mathrm{d}\omega = \left(\frac{\partial g}{\partial z} - \frac{\partial f}{\partial \bar{z}}\right) \mathrm{d}z \wedge \mathrm{d}\bar{z} \tag{2.77}$$

from which it is easy to see that an arbitrary holomorphic differential is closed. The period of a closed differential ω along any closed oriented contour (cycle) γ on X, is defined as $\oint_\gamma \omega$.

The period of a given closed differential depends only on the homology class of the cycle γ. This can be seen by noting that the difference of two homologous curves is a boundary of some region, and we get from Stokes formula

$$\oint_{\gamma - \tilde{\gamma} = \partial\Omega} = \int\int_{\Omega} \mathrm{d}\omega = 0 \ . \tag{2.78}$$

On a surface of genus $g \geq 1$, any cycle γ is homologous to a linear combination of elements of the canonical basis $\{a_1, \ldots, a_g, b_1, \ldots, b_g\}$ of homology cycles with integer coefficients. Further, the pairwise intersection numbers of the $\{a_i, b_i\}$ have the form

$$a_i \circ a_j = b_i \circ b_j = 0, \quad a_i \circ b_j = \delta_{ij}, \qquad i, j = 1, \ldots, g \ . \tag{2.79}$$

The periods of ω with respect to the $\{a_i, b_i\}$ are denoted

$$A_i = \oint_{a_i} \omega, \qquad B_i = \oint_{b_i} \omega \ . \tag{2.80}$$

Riemann's Bilinear Relations: Let X be a Riemann surface of genus g with a canonical basis $\{a_i, b_i\}, i = 1, \ldots, g$ and let ω and ω' be two closed differentials on X with periods A_i, B_i and $A'_i, B'_i, i = 1, \ldots, g$ respectively. Then

$$\int_X \omega \wedge \omega' = \sum_{j=1}^{g} \left(A_j B'_j - A'_j B_j \right), \tag{2.81}$$

where the wedge product of two 1-forms $\omega_1 = p_1 \mathrm{d}z + q_1 \mathrm{d}\bar{z}$ and $\omega_2 = p_2 \mathrm{d}z + q_2 \mathrm{d}\bar{z}$ is given by the 2-form

$$\omega_1 \wedge \omega_2 = (p_1 q_2 - p_2 q_1)\, \mathrm{d}z \wedge \mathrm{d}\bar{z} \ . \tag{2.82}$$

Corresponding to a canonical basis of homology, one can also have a dual basis of holomorphic differentials as

$$\oint_{a_i} \omega_k = 2\pi i \delta_{jk} \ . \tag{2.83}$$

Finally, one also defines a meromorphic 1-form in the same spirit with the holomorphic function $f(z)$ replaced by a meromorphic one.

Definition 11 *A differential form ω on X is called a meromorphic differential if it can be expressed locally in the form*

$$\omega = f(z)\mathrm{d}z \tag{2.84}$$

where f is a meromorphic function of the local parameter z.

We will use these ideas in the next two chapters to understand the low energy effective action of Seiberg-Witten theory.

Chapter 3

Seiberg-Witten Theory

3.1 Introduction

In this chapter we discuss the physical theory, known now as Seiberg-Witten theory, that is characterized by two important symmetry principles – supersymmetry and gauge symmetry. The precise formulation of the idea of supersymmetry and gauge symmetry is lengthy, and we will restrict ourselves to a brief and intuitive introduction of Seiberg-Witten theory. The main focus of the chapter is to understand the appearance of elliptic curves in the solution of the theory, and determine the Seiberg-Witten elliptic curve from monodromy arguments.

We will begin by briefly discussing the symmetry principles that characterize the theory. We will then introduce the objects that play an important role in the theory, and discuss how the objects relate to each other. We will then understand how to solve the theory, and how the data of an elliptic curve emerges naturally from this. The main references for this chapter are [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23].

3.2 Symmetries

The theory we are looking at is defined by the two symmetry principles that underly it – supersymmetry and gauge symmetry. A quantum field theory is described in terms of fields, and the possible interactions among these fields. The physical picture of a field is as a physical entity that 'lives' in space-time, and depends on the coordinates on space-time. Excitations of these fields represent physical particles, carrying energy, momentum, and

having intrinsic degrees of freedom, such as spin. These are the particles that make up the Universe and are observed in particle accelerators.

A field is called bosonic or fermionic depending upon the statistics obeyed by it. Roughly, a bosonic field, $\phi(x)$, obeys the commutation relations[1] $[\phi(x), \phi(y)] = 0$, while a fermionic field, $\psi(x)$, obeys the anti-commutation relations $\{\psi(x), \psi(y)\} = 0$, where x and y are points in space-time.

The basic premise of supersymmetry is the existence of a symmetry of space-time, implemented by an operator Q, called a supersymmetry operator, that acts on the fields, that exchanges bosonic and fermionic fields

$$\begin{aligned} Q\,|\,boson\rangle &= |\,fermion\rangle \\ Q\,|\,fermion\rangle &= |\,boson\rangle. \end{aligned} \tag{3.1}$$

The operator Q acting on a bosonic field generates a specific fermionic field associated to it (and similarly for the application of Q on a fermionic field). Fields related by the action of the supersymmetry operator are known as *superpartners*. Note that the operator Q is idempotent, so that applying it twice on any field gives back that field. Therefore, in any supersymmetric theory, each bosonic field is paired with a corresponding fermionic field, and each particle is paired with an associated supersymmetric partner. This requirement imposes severe restrictions on the possible fields and combinations of fields that can occur in the theory, as well as the interactions allowed between them.

It is also possible to consider the action of more than one supersymmetry operator, for example, Q^1, Q^2, on the fields. This restricts the possibilities of fields and interactions even more than having just one supersymmetry operator. The theory we consider has two supersymmetries acting on it, and is said to have $\mathcal{N} = 2$ supersymmetry. Going ahead, we

[1] Considering the fields $\phi(x)$ as local operators, the product of any two local fields located at the points x and y can be expanded in some basis of local operators as $\phi_a(x)\phi_b(x) \sim \sum_c C^c_{ab}(x-y)\phi_c(y)$ where a, b, c are labels for the local fields and C^c_{ab} are co-efficients that depend on the theory. This product is known as the operator product expansion (OPE).

will note, from time to time, the restrictions $\mathcal{N} = 2$ supersymmetry places on the form of the theory.

The other underlying principle, of gauge invariance is a redundancy in the description of the theory. Fields and interactions that describe the theory are identified up to a group of transformations, known as gauge transformations. The set of these gauge transformations forms a Lie group, and the theory is said to have the corresponding gauge symmetry. The theory we are studying has a SU(2) gauge symmetry. Therefore, the physical theory we are considering is given by a $\mathcal{N} = 2$ supersymmetric SU(2) gauge theory. Further, we restrict ourselves to studying the theory in a limit where excitations with very high energies are integrated out of the field equations. This gives us a low-energy effective theory which does not have any arbitrarily high excitations. It is this theory that we study. We next consider the main objects that constitute the theory.

3.3 Objects of Interest

We will introduce the objects that play an important role in the theory, and where they fit in the description of the theory. We will discuss what constraints apply on them and how the objects are related to one another. Finally, we see what role they play in describing the solution of the theory.

3.3.1 The Prepotential

The first object of interest is the prepotential of the $\mathcal{N} = 2$ supersymmetric SU(2) gauge theory, denoted $\mathcal{F}$. The theory (that we are interested in[2]) is completely determined by an exact determination of the prepotential. This is a consequence of the constraints imposed by requiring $\mathcal{N} = 2$ supersymmetry.

The prepotential is a function of the supersymmetric fields in the theory. The fundamental property that characterizes the $\mathcal{N} = 2$ prepotential is that it depends holomorphically

[2]It is called the Coulomb phase of the theory and is closely related to the theory of electro-magnetism.

on these fields (that is, it depends on a field ϕ, but not on its complex conjugate). When one considers studying the theory in some simplified form, one almost always considers simplifications that retain the holomorphicity property which implies $\mathcal{N} = 2$ supersymmetry.

The expression for the prepotential can be conveniently divided into two parts depending upon the physical phenomenon from which the contribution arises. The sectors are broadly referred to as the perturbative and the non-perturbative sectors and we may write

$$\mathcal{F} = \mathcal{F}_{\text{pert}} + \mathcal{F}_{\text{non-pert}}, \tag{3.2}$$

where $\mathcal{F}_{\text{pert}}$ and $\mathcal{F}_{\text{non-pert}}$ are the perturbative and non-perturbative contributions, respectively. The idea behind each sector can be summarised as follows.

The Perturbative Theory

In quantum field theory, perturbation theory is a set of approximation schemes by which a complicated quantum system is described by successive improvements around a simpler one which gives the leading order description. The approximations are organized in increasing powers of a positive real number $\alpha \ll 1$, known as the *coupling constant*. The physical processes are conveniently represented in terms of Feynman diagrams. Feynman diagrams give a power series in the coupling constant α that organizes the perturbative expansion as an expansion in the number of loops in the Feynman diagram representation. A Feynman diagram with no loops is called a *tree diagram* and forms the leading order contribution. The next order in perturbation theory is given by diagrams with one loop. A Feynman diagram with n loops in it is called an n-loop diagram.

A unique feature in our theory that considerably simplifies the problem is that requiring $\mathcal{N} = 2$ supersymmetry constrains the perturbative contributions to the prepotential to only the tree level and one-loop Feynman diagrams. The form of the perturbative part of the prepotential at one-loop was derived in [12]. There are no further contributions to the prepotential from higher loop terms.

The Non-Perturbative Theory

Gauge theories might also contain a class of allowed field configurations known as instantons. These are represented by terms that cannot be obtained from perturbation theory. However, in considering the quantum theory, they appear as the leading quantum corrections to the classical behavior of a system. These corrections have an intrinsically non-perturbative nature, and cannot be represented in terms of Feynman diagrams. In fact they are negligible in the perturbative regime. The non-perturbative contributions are organized as a power series in an integer k of the form $c_k\, a^2\, (\Lambda/a)^{4k}$, where c_k are real coefficients that have to be determined, a is a complex parameter, that we discuss below, and Λ is a real constant that is fixed in the theory. The integer k is referred to as the instanton number. The one-instanton correction term was derived in [12] and shown to be non-zero. Direct calculation of the coefficients, c_k, even to the first few orders, is a challenge. The calculation of the full set of these corrections is the achievement of Seiberg and Witten's work.

Expression for the Prepotential

The first order of business in solving the theory is to go as far as we can in analyzing the perturbative regime, and determining the prepotential when perturbative effects form the dominant contribution. The perturbative theory can be simplified to the point where we can reflect the contributions in powers of one complex parameter, denoted usually by a. After this simplification, the prepotential $\mathcal{F}$ is a holomorphic function of a. The validity of the perturbative analysis is determined by the magnitude of $|a|$, much like how the approximation to which a Taylor expansion of a function holds is determined by the magnitude of an infinitesimal parameter. We will see below that the coupling constant is proportional to inverse of $\ln(|a|)$, so that when $|a| \to \infty$, the perturbative contributions dominate the non-perturbative ones and the prepotential is almost completely determined by them. This is usually referred to as the *classical* regime in the literature. As $|a|$ decreases

from ∞, the contributions from non-perturbative effects to the prepotential start becoming important. When $a \to 0$, perturbative effects are far dominated by non-perturbative effects and perturbative analysis helps us little in solving the theory.

Thus, constraints from $\mathcal{N} = 2$ supersymmetry imply that the prepotential must be a holomorphic function of the parameter a, and that it has only tree level and one-loop contributions. In the region $|a| \to \infty$, the tree level and one-loop contribution to the prepotential are given by

$$\mathcal{F}(a) \underset{|a|\to\infty}{\sim} \frac{1}{2}\tau_{cl}\, a^2 + \frac{i}{2\pi} a^2 \ln \frac{a^2}{\Lambda^2}, \tag{3.3}$$

where the first term is the tree level contribution and the second term is the one-loop contribution. Here the complex number τ_{cl} is the classical contribution to the gauge coupling constant τ that will be introduced below. Putting everything together, the prepotential including both perturbative and non-perturbative contributions is of the form

$$\mathcal{F} = \frac{1}{2}\tau_{cl}\, a^2 + \frac{i}{2\pi} a^2 \ln \frac{a^2}{\Lambda^2} + \sum_{k=1}^{\infty} c_k\, a^2 \left(\frac{\Lambda}{a}\right)^{4k}, \tag{3.4}$$

where the kth term in the series arises as a contribution of k instantons. One can rescale $\Lambda^2 \mapsto e^{i\pi\tau_{cl}}\Lambda^2$ and absorb the first term into the constant Λ. From here on we write the expression for the prepotential with the tree level term included into the constant Λ which is set to 1. Thus we have the following expression for the pre-potential,

$$\mathcal{F} = \frac{i}{2\pi} a^2 \ln a^2 + \sum_{k=1}^{\infty} c_k\, a^2 \left(\frac{1}{a}\right)^{4k}, \tag{3.5}$$

3.3.2 The Moduli Space

The moduli space, $\mathcal{M}$, of the theory is the parameter space characterizing physically inequivalent theories. This notion is similar to moduli spaces that characterize mathematical objects – curves, tori, etc. An exact determination of this space lies at the heart of solving the theory. In the case of $\mathcal{N} = 2$ supersymmetric SU(2) gauge theories, one complex parameter, denoted u, characterizes inequivalent theories, and the moduli space $\mathcal{M}$ will often

be referred to as 'the u-plane'. We may think of this space as being locally the complex plane.

When $|a| \to \infty$, perturbative contributions determine the prepotential, and hence the theory is determined by a. There is a $\mathbb{Z}_2$ symmetry (that is an action of a subgroup of the SU(2) gauge symmetry mentioned earlier) that implies that theories with parameters a and $-a$ are equivalent. Thus, when a and u are both very large, a good coordinate for the moduli space (which does not distinguish theories related by the $\mathbb{Z}_2$ symmetry) is $u \sim \frac{1}{2}a^2$. The factor of $\frac{1}{2}$ has been included for convenience. In general, we would have the relation

$$u = \frac{1}{2}a^2 + \text{non-perturbative corrections}, \tag{3.6}$$

where the non-perturbative corrections become important when $|a| \ll \infty$. A priori it is not clear how to determine u when $|a|$ is not very large and we have to take non-perturbative corrections into account. Further, $\mathcal{N} = 2$ supersymmetry leads to a $\mathbb{Z}_2$ symmetry on the u plane relating configurations corresponding to u and $-u$.

3.3.3 The Gauge Coupling

Let us introduce the so-called complexified gauge coupling τ. It is a complex parameter whose imaginary part is written as $4\pi/\alpha^2$, where α is the coupling constant that appears in Feynman diagrams. The real part of τ plays a role in certain calculations but will not be important in the present work. The classical τ_{cl} that appears in the first term of the prepotential is the value of τ when no quantum effects are included. It is related to the prepotential as

$$\tau(a) = \frac{\partial^2 \mathcal{F}}{\partial a^2}. \tag{3.7}$$

Taking the large u limit of the expression (3.5) and differentiating it twice with respect to a, we get

$$\tau(a) \sim \frac{i}{\pi}\left(\ln a^2 + 3\right) \qquad \text{as } |a| \to \infty\ . \tag{3.8}$$

As mentioned previously, the imaginary part of $\tau(a)$ is $4\pi/\alpha^2$. Comparing with the expression above, we see that the coupling constant α goes as the inverse of $\ln(|a|)$ as mentioned before.

The logarithm appearing at one-loop implies $\tau(a)$ is a multivalued function of $a^2 \sim 2u$. The imaginary part $\text{Im}\ \tau(a) \sim \frac{1}{\pi}\ln|a|^2$, is positive and single valued for $a^2 \to \infty$.

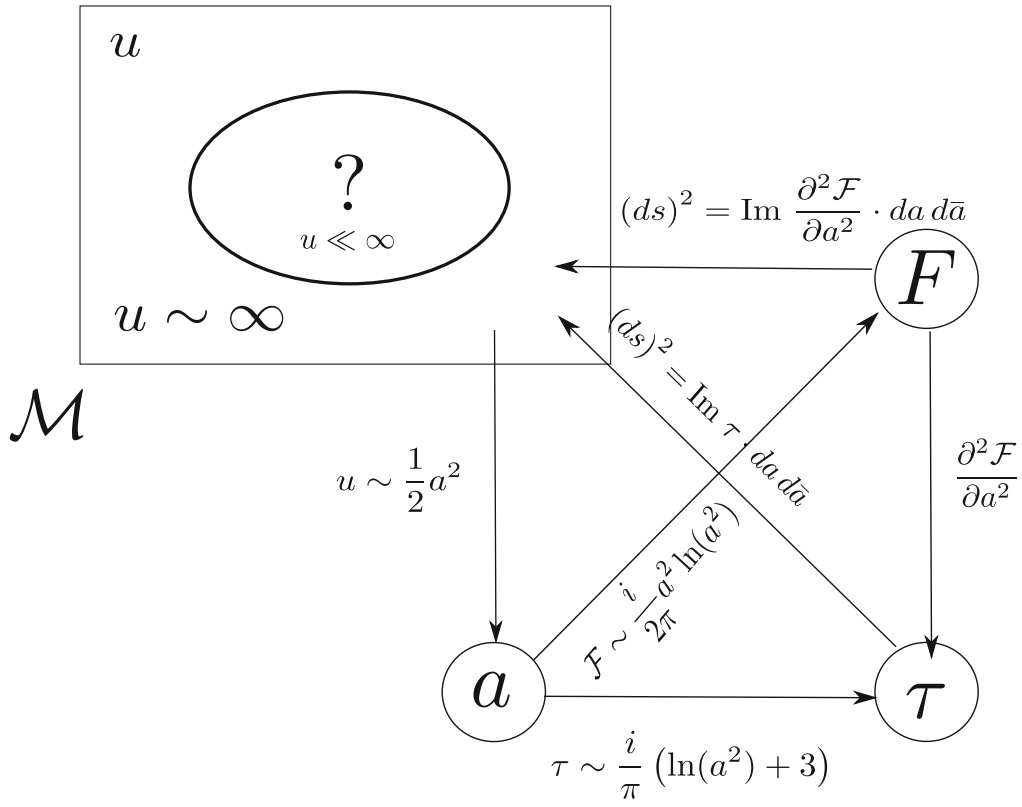

Figure 3.1: The relation between the prepotential, gauge coupling and moduli space for $|a| \to \infty$

Further, from physical considerations, the gauge coupling is related to the moduli space as follows. The metric on the moduli space is given by the imaginary part of the gauge coupling, which locally can be formalized in terms of a as follows:

$$(\mathrm{d}s)^2 = \text{Im}\ \tau(a)\,\mathrm{d}a\mathrm{d}\bar{a} = \text{Im}\ \partial^2\mathcal{F}/\partial a^2\,\mathrm{d}a\mathrm{d}\bar{a}, \tag{3.9}$$

where $\bar{a}$ denotes the complex conjugate of a. The data of the relation between the prepotential, the gauge coupling and the large u region of the moduli space, when $|a| \to \infty$ is shown in the figure 3.1. In the region where $|u| \ll \infty$, however, we do not know how they are related to each other.

3.4 Solving the Theory

The solution to the model lies in determining the prepotential (3.4) exactly including the instanton contributions for all k. So far, we only have the asymptotic expression for $\mathcal{F}(a)$ in (3.3). However, we also have the following constraints on the prepotential:

(1) $\mathcal{F}$ is a holomorphic function of $a \in \mathbb{C}$.

(2) The function $\tau(a) = \frac{\partial^2 \mathcal{F}}{\partial a^2}$ obeys Im $\tau(a) > 0$ since this function is the metric on the moduli space $\mathcal{M}$.

From (3.8) we see that τ is not a single valued function of a. Thus, we are led to the conclusion that the above description of the metric must be valid only locally and there must be a different local parameter that describes τ in the region where a is not the right local parameter.

Changing variables from a to some other local parameter has to, nevertheless, leaves the metric in the form similar to (3.9). Seiberg and Witten proposed rewriting the metric by introducing a new variable given by $a_D = \partial\mathcal{F}/\partial a$, effectively trading $\mathcal{F}$ for a_D and a. With this definition, the metric can be written in the form (3.9) in terms of a_D as

$$(\mathrm{d}s)^2 = \text{Im } \mathrm{d}a_D \, \mathrm{d}\bar{a} = -\frac{i}{2}(\mathrm{d}a_D \, \mathrm{d}\bar{a} - \mathrm{d}a \, \mathrm{d}\bar{a}_D), \tag{3.10}$$

where $\bar{a}_D$ is the complex conjugate of a_D. This formula is symmetric under $(a, a_D) \leftrightarrow (-a_D, a)$, so that if we use a_D as the local parameter the metric will be of the same general form (3.9), but given by a different harmonic function, Im $\tau_D(a_D)$ of a_D. Let us quantify

this statement.

Consider the moduli space $\mathcal{M}$, with some choice of local parameter, ζ. Introduce a two dimensional complex space $X \cong \mathbb{C}^2$ with coordinates (a_D, a). Endow X with the symplectic form $\omega = \text{Im } \mathrm{d}a_D \wedge \mathrm{d}\bar{a}$. The functions $(a_D(\zeta), a(\zeta))$ give a map from $\mathcal{M}$ to X. The metric on $\mathcal{M}$ can be written as

$$(\mathrm{d}s)^2 = \text{Im } \frac{\mathrm{d}a_D}{\mathrm{d}\zeta}\frac{\mathrm{d}\bar{a}}{\mathrm{d}\bar{\zeta}}\,\mathrm{d}\zeta\mathrm{d}\bar{\zeta} = -\frac{i}{2}\left(\frac{\mathrm{d}a_D}{\mathrm{d}\zeta}\frac{\mathrm{d}\bar{a}}{\mathrm{d}\bar{\zeta}} - \frac{\mathrm{d}a}{\mathrm{d}\zeta}\frac{\mathrm{d}\bar{a}_D}{\mathrm{d}\bar{\zeta}}\right)\mathrm{d}\zeta\mathrm{d}\bar{\zeta}. \tag{3.11}$$

The above formula is valid for an arbitrary local parameter ζ on $\mathcal{M}$. Our original description (3.9) corresponds to the choice $\zeta = a$

$$(\mathrm{d}s)^2 = \text{Im } \frac{\mathrm{d}a_D}{\mathrm{d}a}\frac{\mathrm{d}\bar{a}}{\mathrm{d}\bar{a}}\,\mathrm{d}a\mathrm{d}\bar{a} = \text{Im } \frac{\mathrm{d}a_D}{\mathrm{d}a}\,\mathrm{d}a\mathrm{d}\bar{a} = \text{Im } \tau(a)\mathrm{d}a\mathrm{d}\bar{a}, \tag{3.12}$$

since $\dfrac{\mathrm{d}a_D}{\mathrm{d}a} = \dfrac{\mathrm{d}}{\mathrm{d}a}\left(\partial\mathcal{F}/\partial a\right) = \tau(a)$. When expressed this way, we see that the general form of the metric can be preserved by a class of transformations acting on a_D and a. To see the set of possible transformations that do so, we arrange the coordinates into a column vector $v = \begin{pmatrix} a_D(u) \\ a(u) \end{pmatrix}$, and rewrite the metric as

$$(\mathrm{d}s)^2 = \frac{i}{2}\frac{\mathrm{d}v^*}{\mathrm{d}\zeta}\,J\,\frac{\mathrm{d}v}{\mathrm{d}\zeta}\,\mathrm{d}\zeta\mathrm{d}\bar{\zeta}, \qquad \text{with } J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{3.13}$$

It is now easy to see that any transformation of the form $v \mapsto Mv + c$ with $M^*JM = J$ leaves the metric invariant. Let $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, then the condition on M gives

$$\begin{pmatrix} \bar{a} & \bar{c} \\ \bar{b} & \bar{d} \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$
$$\begin{pmatrix} \bar{a}c - a\bar{c} & \bar{a}d - b\bar{c} \\ \bar{b}c - a\bar{d} & \bar{b}d - b\bar{d} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{3.14}$$

This implies, we must have $a, b, c, d \in \mathbb{R}$ and that $ad - bc = 1$. Thus, the set of matrices M satisfying $M^\dagger JM = J$ form the group $\text{Sp}(2, \mathbb{R})$ or equivalently $\text{SL}(2, \mathbb{R})$. Any 2×2 matrix $M \in \text{SL}(2, \mathbb{R})$ acting on $\begin{pmatrix} a_D \\ a \end{pmatrix}$ preserves the form (3.11). It turns out, for the physics to be consistent, the matrix M must be in $\text{SL}(2, \mathbb{Z})$, and the constant $c = 0$. We take this to

be the case from here on. The group $\mathrm{SL}(2,\mathbb{Z})$ is the group, under matrix multiplication, of 2×2 matrices with integer entries having unit determinant:

$$\mathrm{SL}(2,\mathbb{Z}) = \left\{ M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mid a,b,c,d \in \mathbb{Z},\ \det M = 1 \right\} . \tag{3.15}$$

It is generated by the elements

$$S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad \text{and} \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} . \tag{3.16}$$

The element S acting on $v = \begin{pmatrix} a_D \\ a \end{pmatrix}$ gives the transformation

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} a_D \\ a \end{pmatrix} = \begin{pmatrix} a \\ -a_D \end{pmatrix} \tag{3.17}$$

which exchanges $a \leftrightarrow -a_D$. The corresponding gauge coupling $\tau_D(a_D)$ is related to $\tau(a)$ by

$$\tau_D(a_D) = -\frac{1}{\tau(a)} . \tag{3.18}$$

Thus, when $\tau(a)$ approaches zero, we can use $\tau_D(a_D)$ to describe the metric on the moduli space by simultaneously changing from a to a_D. Note that we are still describing the same theory (we are at the same point on the moduli space $\mathcal{M}$), but using a different description $((a_D, \tau_D(a_D))$. Of course, the physical processes in terms of which we previously understood the theory are also changed. That is, the fields and interaction between fields in terms of which we formed our perturbation series are replaced by fields corresponding to the description with a_D and $\tau_D(a_D)$.

This phenomenon is known as '<u>duality</u>' where we can map one description of a theory to another. In our case, we are mapping a description that does not have a perturbative expansion (in terms of a) to another description which can be described perturbatively (in a_D). This is known as S-duality.

3.4.1 Monodromies on the moduli space

Let us look at a_D in the region of the moduli space we had previously considered, i.e. large $|u|$ region of the moduli space. Here the prepotential $\mathcal{F}$ was given by

$$\mathcal{F}(a) \sim \frac{i}{2\pi} a^2 \ln a^2 \ . \tag{3.19}$$

Thus, in the limit of large a, we have

$$a_D = \frac{\partial \mathcal{F}}{\partial a} \approx \frac{2ia}{\pi} \ln a + \frac{ia}{\pi} \ . \tag{3.20}$$

We see that a_D is not a single-valued function of a for large a. Thus, $\begin{pmatrix} a_D \\ a \end{pmatrix}$ has a monodromy in the u-plane. To determine this monodromy, recall that when u and a are large, we have $u \sim \frac{1}{2}a^2$. Going around a closed counterclockwise loop in the u plane around $u = \infty$ we have $u \mapsto e^{-2\pi i} u$, or $\ln u \mapsto \ln u - 2\pi i$ and hence $a \to e^{-i\pi} a$ and $\ln a \mapsto \ln a - \pi i$. The dual variable a_D transforms as

$$a_D \mapsto \frac{2i(-a)}{\pi} \ln \left(e^{-i\pi} a \right) + \frac{i(-a)}{\pi} - 2a = -a_D - 2a. \tag{3.21}$$

Putting these together, the transformations of the variables of going around a circuit in the u plane at large u is

$$\begin{pmatrix} a_D \\ a \end{pmatrix} \to \begin{pmatrix} -1 & -2 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} a_D \\ a \end{pmatrix}. \tag{3.22}$$

Thus, there is a non-trivial monodromy at infinity given by the monodromy matrix

$$M_\infty = \begin{pmatrix} -1 & -2 \\ 0 & -1 \end{pmatrix} \tag{3.23}$$

which can be identified as an element of $\mathrm{SL}(2;\mathbb{Z})$. In fact, there are other points in the u-plane that produce monodromies as u goes around them, as we will see in the next section. It turns out that all the monodromy matrices are in the subgroup $\Gamma(2)$ of $\mathrm{SL}(2,\mathbb{Z})$ consisting of matrices congruent to 1 modulo 2.

3.4.2 Monodromy and Elliptic Curves

We have, thus, reformulated our problem into the *monodromy problem* requiring to determine a, a_D and τ from the following data:

(a) The gauge coupling $\tau(a)$ given in terms of a_D and a by $\tau(a) = \partial a_D/\partial a$ and obeying the condition Im $\tau(a) > 0$.

(b) The monodromy associated to going around ∞ in the u plane being given by M_∞.

Generally, such a monodromy problem is not easy to solve. What makes the present problem solvable is the insight that one can associate an elliptic curve to the data above and compute a and a_D. The condition $\tau(a) > 0$ on the gauge coupling is very reminiscent of the condition on the modular parameter of a torus discussed in the previous chapter. Identifying $\tau(a)$ with the modular parameter, and taking the relation $\tau(a) = \partial a_D/\partial a$, one can identify a and a_D as the periods of a differential around the two fundamental cycles of a torus. Of course, the periods are not unique and a change in the basis of fundamental cycles changes the matrix of periods by an SL$(2, \mathbb{Z})$ matrix. This is also consistent with how a_D and a transform. To, however, determine the exact elliptic curve corresponding to the monodromy problem, we have to determine the number and location of other special points (corresponding to monodromies) besides ∞ in the u plane. We turn to this problem now.

Monodromy at Finite u

The monodromy at infinity implies there must be at least one additional singularity somewhere in the u plane. We previously mentioned there is a discrete $\mathbb{Z}_2$ symmetry $u \leftrightarrow -u$ on the u plane which implies the singularities must come in pairs around $u = 0$. The $\mathbb{Z}_2$ symmetry has two fixed points $u = 0$ and $u = \infty$, and since there is a monodromy associated with $u = \infty$, if there were only two singularities in the u-plane, the other singularity would necessarily have to be at $u = 0$. This, however, implies that the monodromy around

$u = 0$ would be the inverse of that around $u = \infty$ since the contour around $u = 0$ can be deformed into the contour around $u = \infty$: $M_0 = M_\infty^{-1}$. Note that a^2 is left invariant by M_∞. Therefore, if M_∞ was the only monodromy matrix, then u, which is equal to $\frac{1}{2}a^2$, would be a good global coordinate, and the metric would be written in the form (3.9) globally with a global harmonic function $\tau(a)$. We already saw that such functions do not exist, therefore we conclude that there are other special points in the u-plane producing non-trivial monodromy matrices.

The next possibility is to take three singularities at $u = \pm u_0$ (consistent with the $\mathbb{Z}_2$ symmetry) for some finite u_0 and at $u = \infty$, with the monodromies around the three singularities related by $M_\infty^{-1} = M_{u_0} M_{-u_0}$. Physical considerations and consistency checks validate this choice. It is seen from physical arguments that the two points $\pm u_0$ correspond to where $a_D = 0$ and $a + a_D = 0$. It is interesting to note that the large u, large a relation $u \sim \dfrac{1}{2}a^2$, would have led one to expect $u = 0$ to, perhaps, be a singular point on the u plane. Instead, non-perturbative corrections (see (3.6)) dictate that the singularities are at $u = \pm u_0$. Let us analyze the monodromies around the singularities.

Consider first the singularity at $u = u_0$ where

$$a_D(u_0) = 0 \ . \tag{3.24}$$

In the region near $u \to \infty$, $a(u)$ is the preferred local variable, whereas near $u = u_0$, $a_D(u)$ is the preferred local variable. Thus, we may write

$$a_D \approx c_0(u - u_0) \tag{3.25}$$

with some constant c_0. The expression for $\tau_D(a_D)$ is found to be

$$\tau_D(a_D) \approx -\frac{i}{\pi} \ln a_D. \tag{3.26}$$

In terms of a and a_D, $\tau_D(a_D)$ is given by the relation $\tau_D = \mathrm{d}a/\mathrm{d}a_D$. Integrating this, we get

$$a \approx a_0 + \frac{i}{\pi} a_D \ln a_D \approx a_0 + \frac{i}{\pi} c_0 (u - u_0) \ln(u - u_0) \tag{3.27}$$

where $a_0 = a(u = u_0)$ is a non-zero constant. When u goes in a counterclockwise direction around a circuit around u_0, $\ln(u - u_0) \mapsto \ln(u - u_0) + 2\pi i$, and we have

$$\begin{pmatrix} a_D \\ a \end{pmatrix} \mapsto \begin{pmatrix} a_D \\ a - 2a_D \end{pmatrix}. \tag{3.28}$$

We see the monodromy matrix of going around a circuit around u_0 counterclockwise to be

$$M_{u_0} = ST^2S^{-1} = \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}. \tag{3.29}$$

From our assumption that there are only three singularities at $u = \infty$ and at $u = \pm u_0$, and using the argument that a contour around $u = \infty$ can be deformed into a contour circling u_0 followed by a contour circling $-u_0$, we get the condition

$$M_\infty^{-1} = M_{u_0} M_{-u_0} \tag{3.30}$$

on the monodromies. Using this, we determine the monodromy of going around the singularity at $u = -u_0$ counter clockwise as

$$M_{-u_0} = M_{u_0}^{-1} M_\infty^{-1} = \begin{pmatrix} -1 & 2 \\ -2 & 3 \end{pmatrix}. \tag{3.31}$$

We see that the monodromy matrices $M_{u_0}, M_{-u_0}, M_\infty^{-1} \in \Gamma(2)$.

Knowing the singularities in the moduli space and the associated monodromies, we are in a position to solve the low-energy effective theory. The idea, following Seiberg and Witten, is to construct a family of elliptic curves with the given monodromies, and use this to determine the two functions $a = a(u)$ and $a_D = a_D(u)$ satisfying the monodromies (3.23), (3.29), and (3.31). The appearance of the elliptic curve is understood as follows. The monodromy transformations of the vector $(a, a_D)^{\mathrm{T}}$ leads to the transformations of τ by $\tau \mapsto (a\tau + b)/(c\tau + d)$ due to the relation $\tau_D = \mathrm{d}a/\mathrm{d}a_D$. Given that the metric must be positive definite, and the monodromy matrices are elements of the subgroup $\Gamma(2)$ of $\mathrm{PSL}(2, \mathbb{Z})$, this suggests to consider the quotient space $\mathbb{H}/\mathrm{PSL}(2, \mathbb{Z})$ of the upper half plane by $\mathrm{PSL}(2, \mathbb{Z})$. We also have three singularities on the u-space at $u = \infty$ and at

$u = -u_0, u_0$, (we take $u_0 = 1$ and $-u_0 = -1$ with no loss of generality). The quotient of the upper half plane by the modular group, is a moduli space for elliptic curves and its appearance here suggests we can interpret the moduli space of our theory as the moduli space of elliptic curves. The family of curves parametrized by $\mathbb{H}/\mathrm{PSL}(2, Z)$ can be described by the equation

$$y^2 = (x-1)(x+1)(x-u), \tag{3.32}$$

with $(x, y) \in \mathbb{C}^2$ and $u \in \mathbb{CP}^1 \setminus \{1, -1, \infty\}$. The modular parameter τ_u of the elliptic curve (3.32) belongs to $\mathbb{H}/\mathrm{PSL}(2, \mathbb{Z})$. The information contained in τ_u is also encoded by the parameter u, which belongs to the Riemann sphere punctured at three points $1, -1, \infty$. It is well known, that the moduli space $\mathbb{H}/\mathrm{PSL}(2, \mathbb{Z})$ has three special (orbifold) points. We thus have a choice to parameterize the family of curves (3.32) by $\tau_u \in \mathbb{H}/\mathrm{PSL}(2, \mathbb{Z})$ or by $u \in \mathbb{CP}^1 \setminus \{1, -1, \infty\}$. Below (and in physics), the choice is made in favour of the parameter u. The modular parameter τ_u, which is required to satisfy $\mathrm{Im}\ \tau_u > 0$, can then be interpreted as the gauge coupling $\tau(a)$ of our $\mathcal{N} = 2$ supersymmetric gauge theory.

The algebraic curve (3.32), varying as u varies, is known as the Seiberg-Witten curve. We consider this curve in detail in the next chapter and show that the monodromy matrices we have derived do indeed arise from this curve. For every u, there is a genus-one Riemann surface $\mathcal{E}_u$ determined by (3.32). From the equation (3.32) we see that y is given as the square root of a polynomial in x, so we can think of the curve as a double cover of the x-plane branched over $-1, 1, \infty$ and u, and the curve becomes singular precisely when two of the branch points coincide, i.e. when $u = -1, 1$ or ∞. All these ideas are made more precise in the next chapter where we consider the Seiberg-Witten curve and derive the monodromy associated to it in detail.

Chapter 4

Seiberg-Witten Solution

4.1 Introduction

In the previous two chapters we had a brief introduction to Riemann surfaces and the appearance of an elliptic curve in the solution of the $\mathcal{N} = 2$ supersymmetric SU(2) gauge theory. We had monodromy associated to singularities in the u-plane, and the solution of the physical theory given in terms of two functions $a(u)$ and $a_D(u)$. Seiberg and Witten constructed the solutions $a(u)$ and $a_D(u)$ of the theory in terms of an elliptic curve. In this chapter we finally put everything together and construct the solution to the $\mathcal{N} = 2$ gauge theory. We do this using Seiberg and Witten's original approach via the Seiberg-Witten curve. We also note the fact that the solutions can be represented by hypergeometric functions. Finally, we use the solution to compute the instanton numbers for the first few values of instanton numbers k.

4.2 Elliptic Curve

In this section we study the family of Riemann surfaces that we arrived at in the previous chapter

$$\mathcal{E} = \left\{ y^2 = (x-1)(x+1)(x-u) : u \in \mathbb{C}\backslash\{-1,1\} \right\} \tag{4.1}$$

where $x \in \mathbb{CP}^1$. The variable u parametrizes the family of Riemann surfaces and we would like to study this family as we vary the variable u. For each u, the function

$$F : \mathcal{E}_u \to \mathbb{CP}^1, \qquad F(x,y) = x, \tag{4.2}$$

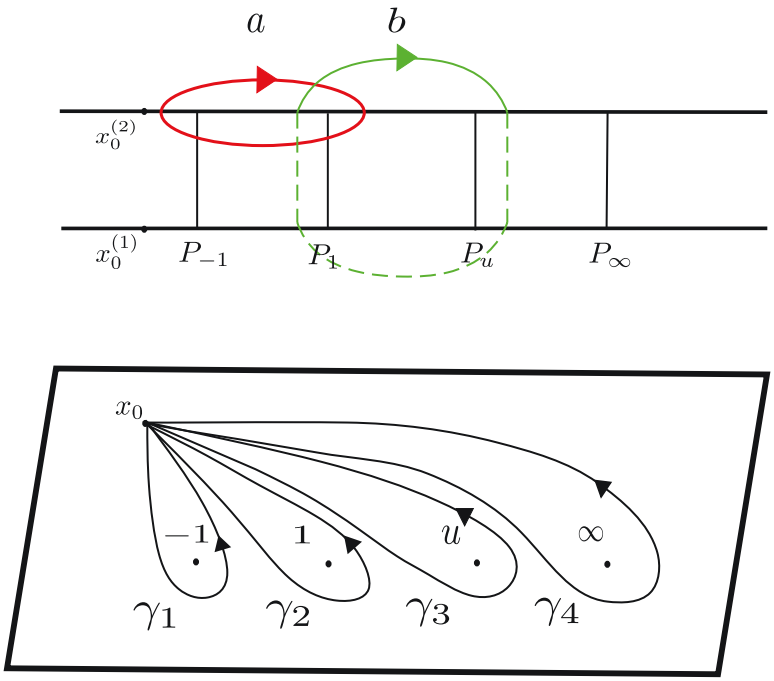

Figure 4.1: Diagram showing $\mathcal{E}_u$ as a covering of the x sphere and the loops $\gamma_i \in \pi_1(\mathbb{CP}^1 \backslash \{-1, 1, u, \infty\}, x_0)$. The vertical lines denote points belonging to both sheets, the ramification points. The corresponding branch points are $x = \pm 1, u, \infty$.

defines a two sheeted ramified covering of the x sphere, where $\mathcal{E}_u$ is the curve

$$\mathcal{E}_u = \{(x, y) \in \mathbb{C}^2 : y^2 = (x-1)(x+1)(x-u)\} \ . \tag{4.3}$$

The compactification of the algebraic curve $\mathcal{E}_u$ is given by the projective curve

$$\hat{\mathcal{E}}_u = \{[\xi, \eta, \zeta] \in \mathbb{CP}^2 : \eta^2 \zeta = (\xi - \zeta)(\xi + \zeta)(\xi - u\zeta)\} \ . \tag{4.4}$$

The projective curve $\hat{\mathcal{E}}_u$ intersects the line at infinity given by $\zeta = 0$ at $[0, 1, 0]$. To describe points in the neighbourhood of the line at infinity we change co-ordinates to (m, n) where

$$m = \frac{\eta\zeta}{\xi^2} = \frac{y}{x^2} \quad n = \frac{\zeta}{\xi} = \frac{1}{x}. \tag{4.5}$$

In these co-ordinates the projective curve is given by

$$m^2 = n(1-n)(1+n)(1-nu) \tag{4.6}$$

whose ramification points are $n = 0, \pm 1, \frac{1}{u}$. In terms of the (y, x) co-ordinates, $n = 0$ corresponds to the point $x = \infty$ and thus the two-sheeted covering is ramified over $x = \pm 1, u, \infty$. Hence, the variable u cannot take the values ± 1 and ∞. As topological spaces, the elliptic curves $\mathcal{E}_u$, for different u, are all isomorphic as long as $u \neq \pm 1, \infty$. However, the Riemann surfaces $\mathcal{E}_u$ and $\mathcal{E}_{u'}$ are not isomorphic as complex manifolds.

Let us understand the topology of the surface $y^2 = (x-1)(x+1)(x-u)$. To each x, there correspond two values of y that differ by a sign. We go from one value of y to the other by analytically continuing $y(x)$ along any closed path going once around one of the roots $1, -1, u$. Considering $y = \sqrt{(x-1)(x+1)(x-u)}$, each of the factors under the square root changes sign when its argument changes by 2π. Thus, the function y is not well defined in the x-sphere. Let us construct a surface on which y will be a well defined single valued function. If we cut the x sphere from -1 to 1, we cannot wind around either ± 1 alone without changing sign. However, if we can choose a path which goes around both -1 and 1, both the factors $\sqrt{x-1}$ and $\sqrt{x+1}$ change sign, and there is no change in y. We also make a cut from u to ∞ to prevent us from winding around all three roots $\pm 1, u$. Thus, either branch of y is single valued in the cut sphere. We take two copies of the cut x sphere and connect them crosswise over the cuts to obtain a two-sheeted Riemann surface on which $y = \sqrt{(x-1)(x+1)(x-u)}$ is single valued (see [24] Chapter 1 for many illustrative figures).

Joining along the two branch cuts makes the covering space into a genus one surface since joining a pair of $\mathbb{CP}^1$ surfaces along two cuts introduces a hole between the two cuts. Thus, the covering space for each value of u is just a complex torus. A loop that goes around one of the two cuts is a homologially non-trivial cycle on the surface. We take the a-cycle to be a loop on one of the copies of x-sphere surrounding the branch cut between -1 and 1. A loop that intersects both cuts would correspond to the b-cycle. The a and b-cycles are shown in red and green in the figure 4.1. The meromorphic function $F : \mathcal{E}_u \to \mathbb{CP}^1$ defined by (4.2) realises a two sheeted covering of $\mathbb{CP}^1$ branched over the points $-1, 1, u, \infty$.

Let us pick two closed contours, ζ_1 and ζ_2 on the genus one surface $\mathcal{E}_u$, normalized such that they intersect with the intersection index one

$$\zeta_1 \circ \zeta_2 = 1 \ . \tag{4.7}$$

The two contours form a canonical basis in the first homology of the surface $\mathcal{E}_u$. Being on a genus one surface, we have a unique (up to scalar multiplication) holomorphic differential

$$\omega = \frac{\mathrm{d}x}{y} = \frac{\mathrm{d}x}{\sqrt{(x-1)(x+1)(x-u)}} \ . \tag{4.8}$$

The periods of the holomorphic differential along the canonical cycles ζ_1 and ζ_2 are

$$A(u) = \oint_{\zeta_1} \frac{\mathrm{d}x}{\sqrt{(x-1)(x+1)(x-u)}}, \quad B(u) = \oint_{\zeta_2} \frac{\mathrm{d}x}{\sqrt{(x-1)(x+1)(x-u)}} \ . \tag{4.9}$$

The ratio of the periods,

$$\tau_u(\mathcal{E}_u, \zeta_1, \zeta_2) = \frac{\oint_{\zeta_2} \omega}{\oint_{\zeta_1} \omega} \tag{4.10}$$

is the modulus of the elliptic curve $\mathcal{E}_u$. As explained in (2.19), τ_u has a positive imaginary part. The periods depend on the choice of the homology basis $\{\zeta_1, \zeta_2\}$, and a different choice of the homology basis gives a different determination of the periods $A(u)$ and $B(u)$. If $\{\delta_1, \delta_2\}$ give a different canonical homology basis, then

$$\begin{aligned} \delta_1 &= a\zeta_1 + b\zeta_2 \\ \delta_2 &= c\zeta_1 + d\zeta_2, \end{aligned} \tag{4.11}$$

where $a, b, c, d \in \mathbb{Z}$ and the matrix

$$T = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{4.12}$$

has unit determinant. The periods in the new basis are given as

$$\begin{aligned} A' &= aA + bB \\ B' &= cA + dB \ . \end{aligned} \tag{4.13}$$

The invariant τ_u transforms to $\tau' = B'/A'$ and is related to τ_u as

$$\tau' = \frac{c + d\tau_u}{a + b\tau_u} \ . \tag{4.14}$$

The periods $A(u)$ and $B(u)$ are holomorphic functions of u, as can be seen by computing $\partial A/\partial \bar{u}$ and $\partial B/\partial \bar{u}$ by differentiating under the integral sign (since the domain of integration is constant).

On any simply connected open set U of $\mathbb{CP}^1 \backslash \{-1, 1, \infty\}$, the periods $A(u), B(u)$ and the ratio τ_u are all single valued functions. On the full domain $\mathbb{CP}^1 \backslash \{-1, 1, \infty\}$, however, they are multivalued. We study the monodromy representation of the fundamental group of the u-sphere punctured at the points $-1, 1, \infty$ associated with the family of curves $\mathcal{E}_u$ next.

4.3 Monodromy Representation

We now study the monodromy representation of the fundamental group of the u-space $\mathbb{CP}^1 \backslash \{-1, 1, \infty\}$. This is a representation

$$\rho : \pi_1(\mathbb{CP}^1 \backslash \{-1, 1, \infty\}) \to \mathrm{GL}(2, \mathbb{C}) \ . \tag{4.15}$$

When u coincides with one of the branch points, as it varies on the domain $\mathbb{CP}^1 \backslash \{-1, 1, \infty\}$, one of the cycles on the torus shrinks to zero size. For example, as u goes to ∞, the a-cycle shrinks to zero size and as $u \to 1$, the b-cycle shrinks to zero size (see figure 4.1). Thus, at $u = 1, -1, \infty$, the curve $\mathcal{E}$ has a vanishing cycle.

4.3.1 Calculating the monodromy matrices

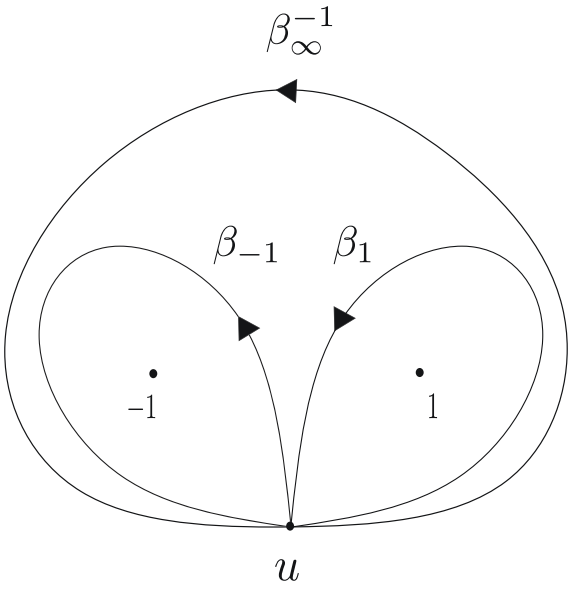

Figure 4.2: The u-plane showing the loops β_1, β_{-1} and β_∞.

As u goes around a loop in the moduli space, the associated genus one surface $\mathcal{E}_u$ varies along with it. Considering a loop around one of the punctures in the u-sphere gives one a transformation for the basis of homology cycles of the curve $\mathcal{E}_u$. We call this transformation the monodromy transformation.

Let us denote loops in the u sphere going in a counterclockwise direction around the points $1, -1$ and ∞ by β_1, β_{-1} and β_∞ respectively (see figure 4.2). As u goes around the singularities 1 and -1, along the cycle $\beta_1, \beta_{-1} \in \pi_1(\mathbb{CP}^1 \setminus \{-1, 1, \infty\}, x_0)$ it leads to the corresponding monodromy matrices

$$M_{-1} = \rho(\beta_{-1}), \quad M_1 = \rho(\beta_1) \ . \tag{4.16}$$

We would like to calculate these monodromy matrices for the family given by (4.1).

Any path $\zeta : [0, 1] \to \mathbb{CP}^1$ can be lifted to a path $\hat{\zeta}$ on $\mathcal{E}_u$ using the branched covering $F : \mathcal{E}_u \to \mathbb{CP}^1$. If $\zeta(0) = x_0$ is a point on $\mathbb{CP}^1$, then there exists a point $x_0' \in \mathcal{E}_u$ and a

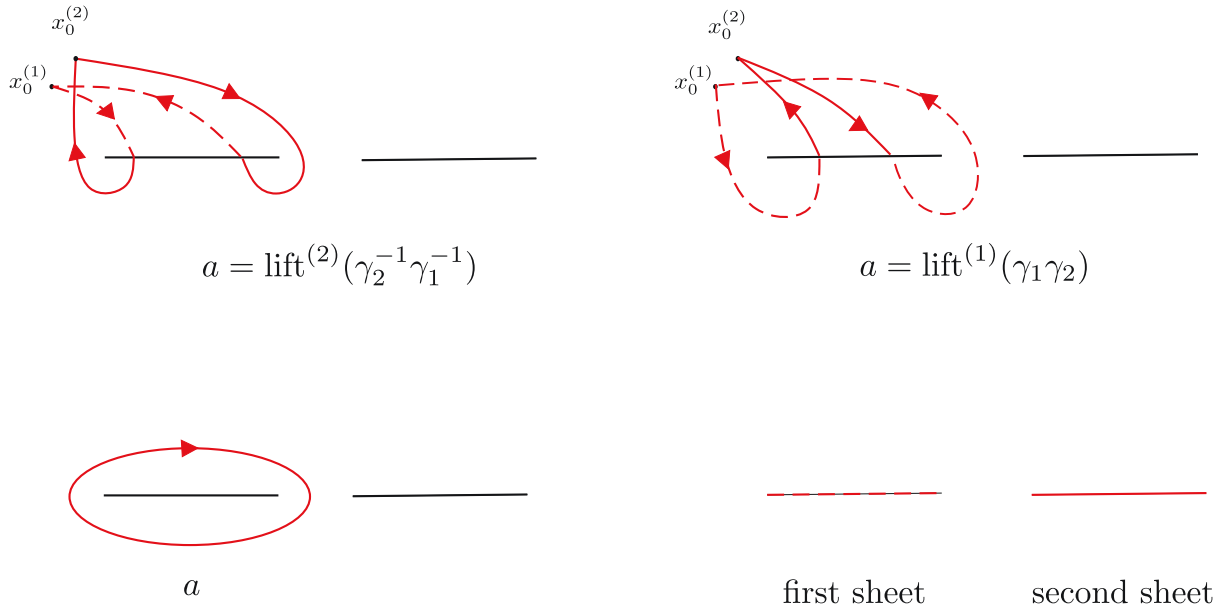

Figure 4.3: Period cycle a expressed in terms of the lifts of loops γ_1 and γ_2. Dotted lines indicate parts of the loop that lie on the first sheet. $a = \text{lift}^{(2)}(\gamma_1^{-1}\gamma_2^{-1}) = \text{lift}^{(1)}(\gamma_1\gamma_2)$

lifted path $\hat{\zeta} \subset \mathcal{E}_u$ such that $\hat{\zeta}(0) = x_0'$ and $F \circ \hat{\zeta} = \zeta$. We denote the lift of the path ζ to the first/second sheet of the covering by $\text{lift}^{(1/2)}(\zeta)$.

To see what happens to the genus one surface $\mathcal{E}_u$ as u goes around the loops β_i, we have to understand how the a and b cycles transform as one goes around each of the β_i's. The a and b cycles on the torus are in turn related to the branch points on the x-plane and given as lifts of the loops γ_i in the figure 4.1. The lifts for the a and b cycles are given as follows. The a-cycle goes around the points 1 and -1 and is therefore given in terms of the lifts of the loops γ_1 and γ_2 as (see figures 4.3 and 4.8)

$$a = \text{lift}^{(2)}\left(\gamma_2^{-1}\gamma_1^{-1}\right) = \text{lift}^{(1)}\left(\gamma_1\gamma_2\right) = \text{lift}^{(1)}\left(\gamma_1^{-1}\gamma_2\right) \ . \tag{4.17}$$

Similarly the b-cycle goes around the points u and 1 and is given in terms of the lifts of the

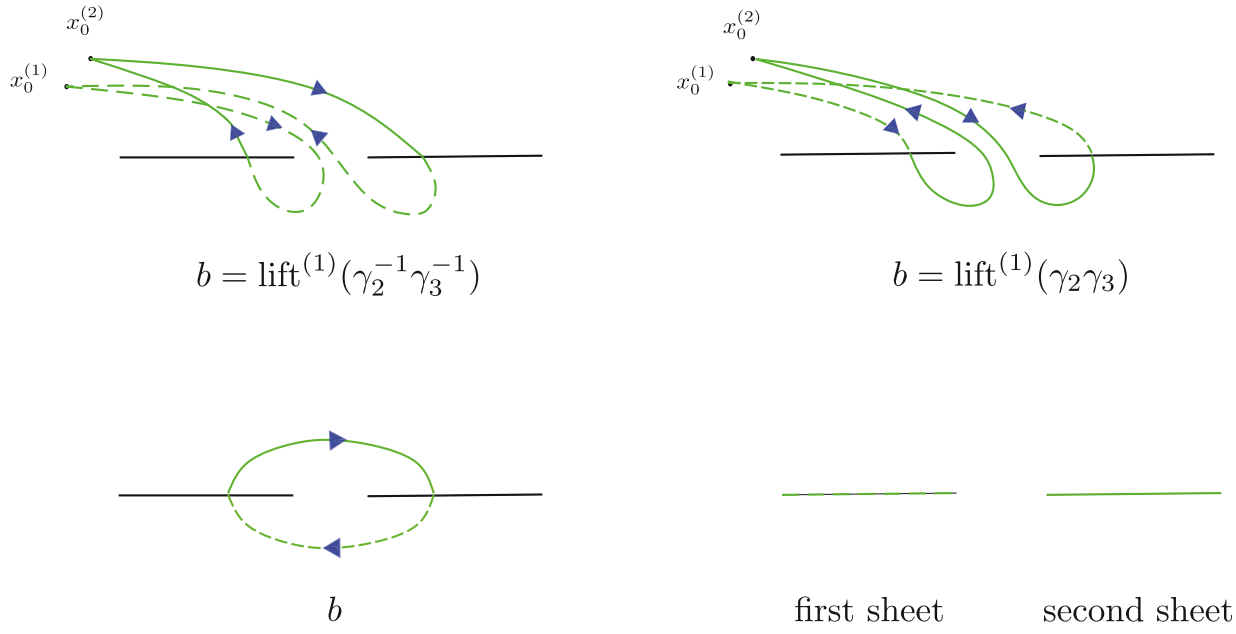

Figure 4.4: Period cycle b expressed in terms of the loops γ_2 and γ_3. $b = \text{lift}^{(1)}(\gamma_2\gamma_3) = \text{lift}^{(2)}(\gamma_2^{-1}\gamma_3^{-1})$

loops γ_2 and γ_3 as (see figure 4.4)

$$b = \text{lift}^{(2)}\left(\gamma_2^{-1}\gamma_3^{-1}\right) = \text{lift}^{(1)}\left(\gamma_2\gamma_3\right) \ . \tag{4.18}$$

We also have (see figures 4.5 and 4.6),

$$-b = \text{lift}^{(2)}(\gamma_3^{-1}\gamma_2^{-1}) = \text{lift}^{(1)}(\gamma_3\gamma_2) = \text{lift}^{(1)}\left(\gamma_3\gamma_2^{-1}\right) \ . \tag{4.19}$$

Now, to see how the a and b-cycles transform we need to know how the loops γ_i in the x sphere transform as the variable u goes along the loops β_i.

Calculating M_{-1}

We want to calculate the action of β_{-1} on the a and b cycles by following the action of β_{-1} on the loops γ_i. In figure 4.7 we see what happens to the loops γ_i as u goes along the loop β_{-1}. We denote by $\tilde{\gamma}_i$ the action $\beta_{-1}(\gamma_i)$ of β_{-1} on the loop γ_i. The loop $\tilde{\gamma}_3$ is marked in

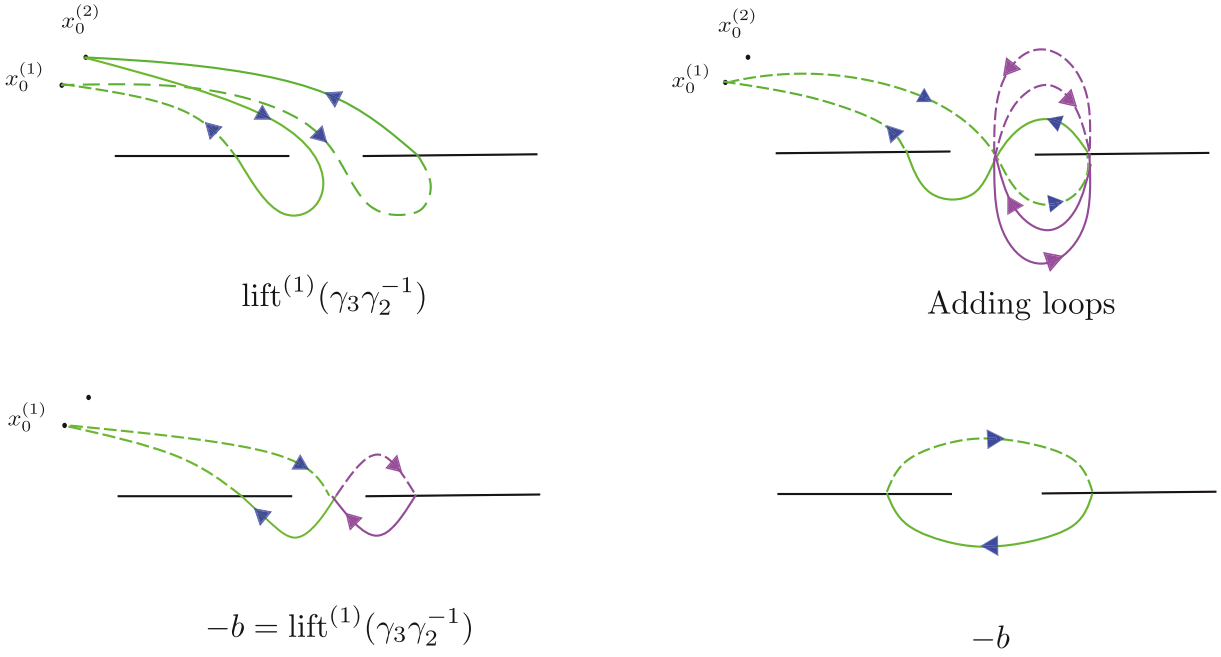

Figure 4.5: $-b = \text{lift}^{(1)}(\gamma_3\gamma_2^{-1})$

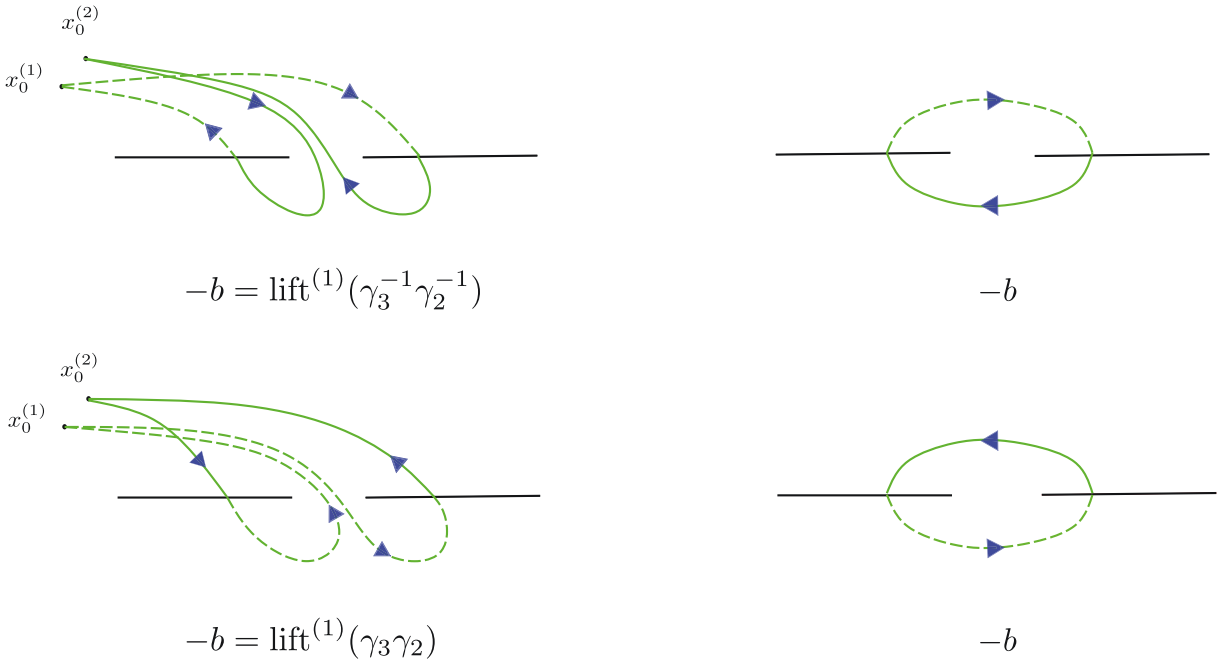

Figure 4.6: Negative of the period cycle b expressed in terms of the loops γ_2 and γ_3. $-b = \text{lift}^{(2)}(\gamma_3^{-1}\gamma_2^{-1}) = \text{lift}^{(1)}(\gamma_3\gamma_2)$

red. From figure 4.7 we see that,

$$\beta_{-1}(\gamma_1) = \gamma_1\,\gamma_2\,\gamma_3\,\gamma_2^{-1}\,\gamma_1\,\gamma_2\,\gamma_3^{-1}\,\gamma_2^{-1}\,\gamma_1^{-1}$$
$$\beta_{-1}(\gamma_2) = \gamma_2$$
$$\beta_{-1}(\gamma_3) = \gamma_2^{-1}\,\gamma_1\,\gamma_2\,\gamma_3\,\gamma_2^{-1}\,\gamma_1^{-1}\,\gamma_2 \quad .$$

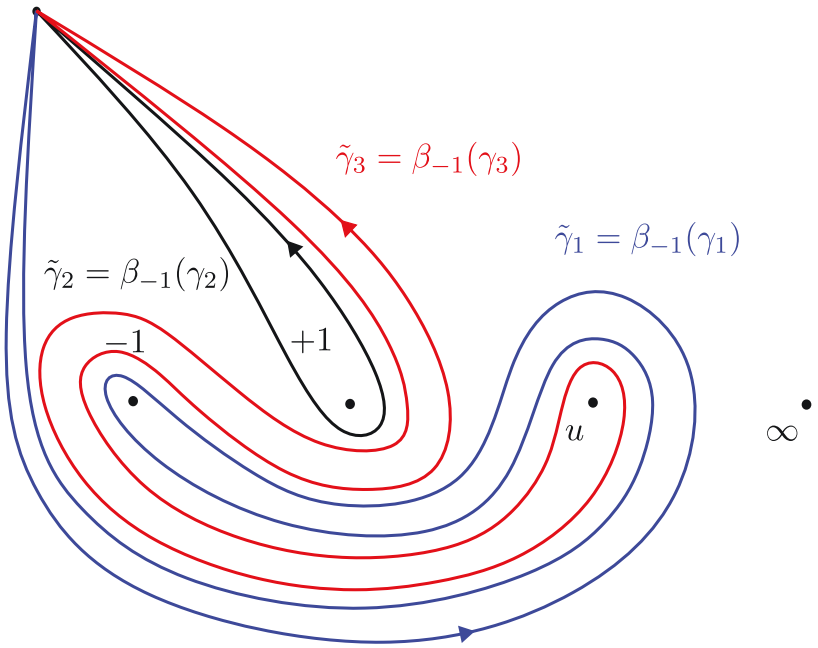

Figure 4.7: The action of β_{-1} on γ_is

Therefore, for the action of β_{-1} on the a cycle we get

$$a = \text{lift}^{(1)}(\gamma_1\,\gamma_2)$$
$$\tilde{a} = \beta_{-1}(a) = \text{lift}^{(1)}\Big(\beta_{-1}(\gamma_1)\beta_{-1}(\gamma_2)\Big)$$
$$\tilde{a} = \text{lift}^{(1)}\Big(\underbrace{\gamma_1\,\gamma_2}_{\text{a}}\,\underbrace{\gamma_3\,\gamma_2^{-1}}_{\text{-b}}\,\underbrace{\gamma_1\,\gamma_2}_{\text{a}}\,\underbrace{\gamma_3^{-1}\,\gamma_2^{-1}}_{\text{-b}}\,\underbrace{\gamma_1^{-1}\,\gamma_2}_{\text{a}}\Big)$$
$$\tilde{a} = -2b + 3a \ .$$

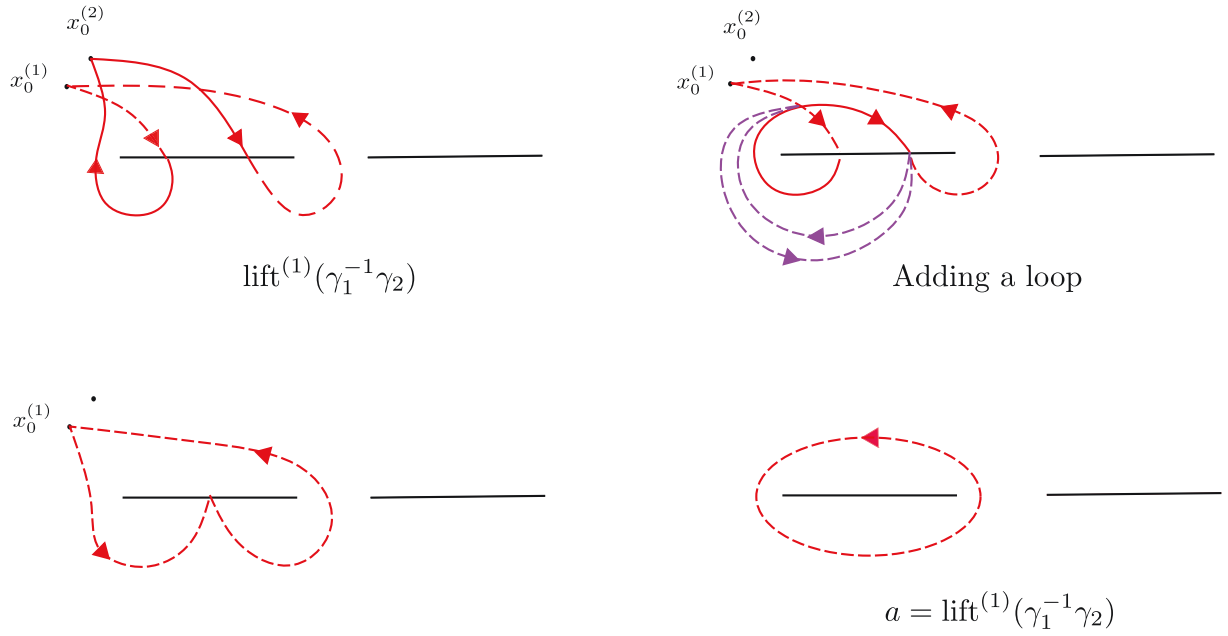

Figure 4.8: $a = \text{lift}^1(\gamma_1^{-1}\gamma_2)$

For the action of β_{-1} on the b cycle we get

$$
\begin{aligned}
b &= \text{lift}^{(1)}(\gamma_2\gamma_3) \\
\tilde{b} &= \beta_{-1}(b) = \text{lift}^{(1)}\Big(\beta_{-1}(\gamma_2)\beta_{-1}(\gamma_3)\Big) \\
\tilde{b} &= \text{lift}^{(1)}\Big(\underbrace{\gamma_2\,\gamma_2^{-1}}_{1}\underbrace{\gamma_1\,\gamma_2}_{\text{a}}\,\underbrace{\gamma_3\gamma_2^{-1}}_{\text{-b}}\underbrace{\gamma_1^{-1}\gamma_2}_{\text{a}}\Big) \\
\tilde{b} &= -b + 2a\ .
\end{aligned}
$$

Thus, the action of β_{-1} on the a and b cycles is given by

$$\tilde{a} = -2b + 3a \tag{4.20}$$

$$\tilde{b} = -b + 2a, \tag{4.21}$$

and the corresponding matrix, M_{-1}, is thus found to be

$$\begin{pmatrix}\tilde{b}\\ \tilde{a}\end{pmatrix} = \begin{pmatrix}-1 & 2\\ -2 & 3\end{pmatrix}\begin{pmatrix}b\\ a\end{pmatrix}; \qquad M_{-1} = \begin{pmatrix}-1 & 2\\ -2 & 3\end{pmatrix}\ . \tag{4.22}$$

4.3.2 Calculating M_1

We can similarly track the action of β_1 on the loops γ_i and find the action of β_1 on the a and b cycles. From figure 4.9 we see the action of β_1 on the loops γ_i, denoted $\tilde{\gamma}_i$ in the figure, to be

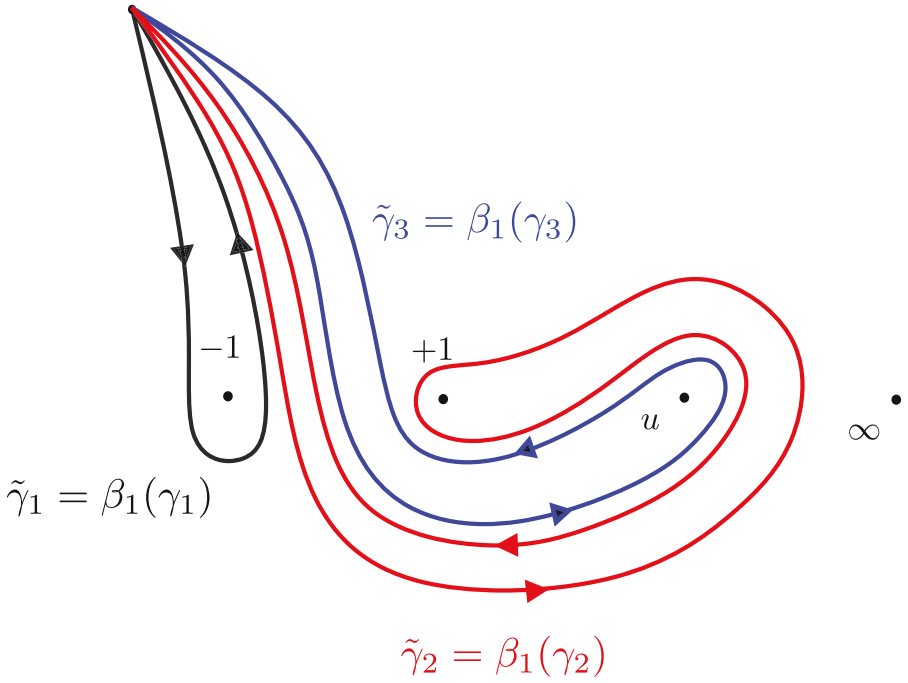

Figure 4.9: The action of β_1 on γ_is

$$\beta_1(\gamma_1) = \gamma_1 \tag{4.23}$$

$$\beta_1(\gamma_2) = \gamma_2\,\gamma_3\,\gamma_2\,\gamma_3^{-1}\,\gamma_2^{-1} \tag{4.24}$$

$$\beta_1(\gamma_3) = \gamma_2\,\gamma_3\,\gamma_2^{-1}\ . \tag{4.25}$$

Therefore, the action of β_1 on the a cycle is

$$\begin{aligned}
a &= \text{lift}^{(1)}(\gamma_1\gamma_2) \\
\tilde{a} &= \beta_1(a) = \text{lift}^{(1)}\Big(\beta_1(\gamma_1)\beta_1(\gamma_2)\Big) \\
\tilde{a} &= \text{lift}^{(1)}\Big(\underbrace{\gamma_1\gamma_2}_{a}\ \underbrace{\gamma_3\,\gamma_2}_{-b}\ \underbrace{\gamma_3^{-1}\,\gamma_2^{-1}}_{-b}\Big) \\
\tilde{a} &= -2b + a\ .
\end{aligned}$$

The action of β_1 on the b cycle is

$$\begin{aligned}
b &= \text{lift}^{(1)}(\gamma_2\gamma_3) \\
\tilde{b} &= \beta_1(b) = \text{lift}^{(1)}\Big(\beta_1(\gamma_2)\beta_1(\gamma_3)\Big) \\
\tilde{b} &= \text{lift}^{(1)}\Big(\gamma_2\,\gamma_3\ \underbrace{\gamma_2\ \overbrace{\gamma_3^{-1}\ \underbrace{\gamma_2^{-1}\gamma_2}\ \gamma_3}\ \gamma_2^{-1}}\Big) \\
\tilde{b} &= \text{lift}^{(1)}(\gamma_2\gamma_3) = b\ .
\end{aligned}$$

Putting the two together, the action of β_1 on the a and b cycles is

$$\tilde{b} = b \tag{4.26}$$

$$\tilde{a} = -2b + a, \tag{4.27}$$

and the monodromy matrix M_1 corresponding to this action on a and b is

$$\begin{pmatrix}\tilde{b}\\ \tilde{a}\end{pmatrix} = \begin{pmatrix}1 & 0\\ -2 & 1\end{pmatrix}\begin{pmatrix}b\\ a\end{pmatrix}; \qquad M_1 = \begin{pmatrix}1 & 0\\ -2 & 1\end{pmatrix}\ . \tag{4.28}$$

Finally, computing M_∞ from the relation $M_\infty^{-1} = M_1 M_{-1}$, we arrive at the three monodromy matrices we derived in the previous chapter:

$$M_{-1} = \begin{pmatrix}-1 & 2\\ -2 & 3\end{pmatrix}, \quad M_1 = \begin{pmatrix}1 & 0\\ -2 & 1\end{pmatrix}, \quad M_\infty^{-1} = \begin{pmatrix}-1 & 2\\ 0 & -1\end{pmatrix}\ . \tag{4.29}$$

We have checked that the elliptic curve defined by $y^2 = (x-1)(x+1)(x-u)$ does indeed have the monodromy structure we require from our gauge theory considerations. We can thus interpret the period matrix of this elliptic curve as the gauge coupling of our gauge theory and construct it as the ratio of the period integrals of the curve.

4.4 Exact solution from elliptic curves

To solve the model we have to determine the holomorphic prepotential from the two holomorphic functions $a(u)$ and $a_D(u)$. Starting with the Seiberg-Witten elliptic curve

$$y^2 = (x-1)(x+1)(x-u) \tag{4.30}$$

we want to determine the functions $a(u)$ and $a_D(u)$. The modulus τ_u of the elliptic curve is defined as the ratio of the periods $\omega(u)$ and $\omega_D(u)$, given as the integrals over the a and b-cycles of the unique holomorphic closed one-form $\frac{\mathrm{d}x}{y}$

$$\omega(u) = \oint_{a-\text{cycle}} \frac{\mathrm{d}x}{y} \qquad \omega_D(u) = \oint_{b-\text{cycle}} \frac{\mathrm{d}x}{y} \tag{4.31}$$

with

$$\frac{\mathrm{d}x}{y} = \frac{\mathrm{d}x}{\sqrt{(x-1)(x+1)(x-u)}}\ . \tag{4.32}$$

A different choice of the a and b-cycles will transform the pair (ω_D, ω) by an $\mathrm{SL}(2,\mathbb{Z})$ matrix and τ_u would be transformed by the standard action of $\mathrm{SL}(2,\mathbb{Z})$ on the upper half plane (see eq. (2.19) in Section 2.2). The modulus τ_u also satisfies the fundamental property

$$\mathrm{Im}(\tau_u) > 0. \tag{4.33}$$

Now, we identify our SU(2) gauge theory with the above curve by identifying the modulus of the torus with the complexified gauge coupling, $\tau(u)$, and the periods ω and ω_D with the u derivatives of a and a_D respectiely:

$$\tau_u = \frac{\omega_D}{\omega} = \frac{\oint_b \frac{\mathrm{d}x}{y}}{\oint_a \frac{\mathrm{d}x}{y}} = \frac{\mathrm{d}a/\mathrm{d}u}{\mathrm{d}a_D/\mathrm{d}u} = \tau(u)\,. \tag{4.34}$$

The condition Im $\tau_u > 0$ therefore implies Im $\tau(u) > 0$, which gives positivity of the metric. Since u is globally defined, a and a_D have the same monodromies as ω and ω_D. Integrating

$$\frac{\mathrm{d}a}{\mathrm{d}u} = \oint_{a-\text{cycle}} \frac{\mathrm{d}x}{y}, \qquad \frac{\mathrm{d}a_D}{\mathrm{d}u} = \oint_{b-\text{cycle}} \frac{\mathrm{d}x}{y} \tag{4.35}$$

on both sides with respect to u gives

$$a = \oint_{a-\text{cycle}} \lambda, \qquad a_D = \oint_{b-\text{cycle}} \lambda, \tag{4.36}$$

with λ given by

$$\lambda = \frac{\sqrt{2}}{2\pi}\frac{(x-u)\mathrm{d}x}{y} = \frac{\sqrt{2}}{2\pi}\frac{(x-u)\mathrm{d}x}{\sqrt{(x^2-1)(x-u)}} = \frac{\sqrt{2}}{2\pi}\mathrm{d}x\sqrt{\frac{x-u}{x^2-1}}, \tag{4.37}$$

with the factor $\sqrt{2}/2\pi$ being dictated by the asymptotic behavior of a_D and a near $u = 1, -1, \infty$. With the a and b cycles on $\mathcal{E}_u$ as described previously, we can express the above integrals as

$$a(u) = \frac{\sqrt{2}}{\pi}\int_{-1}^{1}\frac{dx\sqrt{x-u}}{\sqrt{x^2-1}}, \tag{4.38}$$

and

$$a_D(u) = \frac{\sqrt{2}}{\pi}\int_{1}^{u}\frac{dx\sqrt{x-u}}{\sqrt{x^2-1}}, \tag{4.39}$$

where an additional factor of 2 comes from the fact that the integral is taken from -1 to 1 and then from 1 back to -1. These are the expressions for the two functions $a(u)$ and $a_D(u)$ that we sought to find. We can now check that they have the right asymptotic behavior by studying their behavior near $u = \infty$ and $u = 1$. Recall that the u plane had a $\mathbb{Z}_2$ symmetry between u and $-u$ (see Section 3.3.2). Thus, the behavior near $u = -1$ can be determined from the behavior near $u = 1$.

4.4.1 Behavior of $a(u)$ and $a_D(u)$ for u near ∞

$a(u)$ at ∞:

The behavior of $a(u)$ as $u \to \infty$ is seen to be

$$\begin{aligned} a(u) &= \frac{\sqrt{2}}{\pi}\int_{-1}^{1}\frac{dx\sqrt{x-u}}{\sqrt{x^2-1}} \\ &\approx \frac{\sqrt{2u}}{\pi}\int_{-1}^{1}dx\frac{1}{\sqrt{1-x^2}}\left(1-\frac{x}{2u}-\frac{x^2}{8u^2}+\mathcal{O}\left(\frac{x^3}{u^3}\right)\right) \\ &\approx \frac{\sqrt{2u}}{\pi}\left(\pi-\frac{\pi}{16u^2}+\mathcal{O}\left(\frac{1}{u^3}\right)\right) \approx \sqrt{2u}, \end{aligned} \tag{4.40}$$

which gives us $u = \frac{1}{2}a^2$ as expected.

a_D at ∞:

For $u \to \infty$ the integral for $a_D(u)$ developes a logarithmic divergence near $z = 0$. The behavior of $a_D(u)$ as $u \to \infty$ is seen as follows

$$\lim_{u\to\infty} a_D(u) = \frac{\sqrt{2}}{\pi}\int_1^u \frac{dx\,\sqrt{x-u}}{\sqrt{x^2-1}}$$

$$= \frac{\sqrt{2/\epsilon}}{\pi}\int_\epsilon^1 \frac{dz\,\sqrt{z-1}}{\sqrt{z^2-\epsilon^2}} \qquad \text{with} \quad z = \frac{x}{u}, \quad \epsilon = \frac{1}{u}\ . \tag{4.41}$$

Splitting this integral into two parts, we obtain

$$\lim_{u\to\infty} a_D(u) = \frac{\sqrt{2/\epsilon}}{\pi}\left(\int_\epsilon^A \frac{dz\,\sqrt{z-1}}{\sqrt{z^2-\epsilon^2}} + \int_A^1 \frac{dz\,\sqrt{z-1}}{\sqrt{z^2-\epsilon^2}}\right) \qquad \text{where} \ \ \epsilon \ll A \ll 1\ .$$

The first term, where z is comparable to ϵ, can be expanded as

$$\int_\epsilon^A \frac{dz\,\sqrt{z-1}}{\sqrt{z^2-\epsilon^2}} \simeq i\int_\epsilon^A \frac{dz}{\sqrt{z^2-\epsilon^2}}\Big(1 - \frac{z}{2} + \frac{z^2}{8}\cdots\Big)$$

$$= i\left(\ln\left(z+\sqrt{z^2-\epsilon^2}\right) - \frac{1}{2}\sqrt{z^2-\epsilon^2} + \frac{1}{8}\left(\frac{1}{2}z\sqrt{z^2-\epsilon^2} + \frac{1}{2}\epsilon^2\ln\left(z+\sqrt{z^2-\epsilon^2}\right)\right) + \cdots\right)\Bigg|_\epsilon^A$$

$$\approx i\left(\ln A - \ln\epsilon - \frac{A}{2} + \frac{A^2}{16} + \mathcal{O}(\epsilon)\right) \tag{4.42}$$

while the second term, where $\epsilon \ll z$, can be expanded as

$$\int_A^1 \frac{dz\,\sqrt{z-1}}{\sqrt{z^2-\epsilon^2}} = i\int_A^1 \frac{dz\,\sqrt{1-z}}{\sqrt{z^2-\epsilon^2}}$$

$$\simeq i\int_A^1 \frac{dz\,\sqrt{1-z}}{z}\left(1 + \frac{\epsilon^2}{2z^2} + \frac{3\epsilon^4}{8z^4} + \cdots\right) \qquad \text{expanding } \frac{1}{\sqrt{z^2-\epsilon^2}} \text{ in } \frac{\epsilon}{z}$$

$$= i\int_A^1 dz\left(\frac{\sqrt{1-z}}{z} + \frac{\epsilon^2}{2}\frac{\sqrt{1-z}}{z^3} + \cdots\right)$$

$$= i\int_A^1 \frac{dz}{z}\Big(1 - \frac{z}{2} + \frac{z^2}{8}\cdots\Big) + \mathcal{O}(\epsilon^2)$$

$$\approx -i\left(\ln A - \frac{A}{2} + \frac{A^2}{16} + \mathcal{O}(\epsilon)\right). \tag{4.43}$$

Thus, we see that to leading order

$$\lim_{u\to\infty} a_D(u) \approx \frac{i\sqrt{2u}}{\pi} \ln u, \tag{4.44}$$

which agrees with what we found in (3.20). Next we evaluate the behavior of $a_D(u)$ and $a(u)$ in the region of the moduli space with $u = 1$.

4.4.2 Behavior of $a_D(u)$ and $a(u)$ near $u = 1$

$a_D(u)$ at $u = 1$:

We now derive the asymptotic expansion of $a_D(u)$ when u is near 1. Introducing a new small parameter z such that $u = 1 + 2z$, the expression for $a_D(u)$ can be written as

$$\begin{aligned}
a_D(u) &= \frac{\sqrt{2}}{\pi}\int_1^u dx \frac{\sqrt{x-u}}{\sqrt{x^2-1}} \\
&= \frac{\sqrt{2}i}{\pi}\int_1^{1+2z} dx \frac{\sqrt{1-x+2z}}{\sqrt{x^2-1}} \\
&= \frac{\sqrt{2}i}{\pi}\int_1^{1+2z} dx \frac{\sqrt{1-x+2z}}{\sqrt{x-1}} \frac{1}{\sqrt{2}\sqrt{1+\frac{(x-1)}{2}}} \\
&= \frac{i}{\pi}\int_1^{1+2z} dx \frac{\sqrt{1-x+2z}}{\sqrt{x-1}} \left(1 - \frac{x-1}{4} + \frac{3(x-1)^2}{32}\right. \\
&\qquad\qquad \left. - \frac{5(x-1)^3}{128} + \mathcal{O}((x-1)^4)\right),
\end{aligned} \tag{4.45}$$

where in the last step we have expanded the factor $\sqrt{1+(x-1)/2}$ since $x \sim 1$ and hence $(x-1) \ll 1$. Computing the integrals, we get

$$a_D(z) = i\left(z - \frac{z^2}{8} - \frac{3z^3}{64} + \frac{25z^4}{1024} + \mathcal{O}(z^5)\right) . \tag{4.46}$$

The leading order gives

$$a_D(u) \approx i\frac{(u-1)}{2} . \tag{4.47}$$

$a(u)$ at $u = 1$:

To derive the asymptotic behavior of $a(u)$ near $u = 1$, we write the expression for $a(u)$ as

$$\begin{aligned} a(u) &= \frac{\sqrt{2}}{\pi}\int_{-1}^{1}\frac{dx\,\sqrt{x-u}}{\sqrt{x^2-1}} \\ &= \frac{\sqrt{2}}{\pi}\int_{-1}^{1}dx\frac{\sqrt{1-x+2z}}{\sqrt{1-x^2}} \qquad \text{changing variables } z=(u-1)/2 \\ &= \frac{\sqrt{2}}{\pi}\int_{-1}^{1-Nz}dx\frac{\sqrt{1-x+2z}}{\sqrt{1-x^2}} + \frac{\sqrt{2}}{\pi}\int_{1-Nz}^{1}dx\frac{\sqrt{1-x+2z}}{\sqrt{1-x^2}} \qquad \text{with} \quad 1 \ll N \ll \frac{1}{z}\,. \end{aligned}$$

We evaluate the integrals separately. The first integral can be evaluated as

$$\frac{\sqrt{2}}{\pi}\int_{-1}^{1-Nz}dx\frac{\sqrt{1-x+2z}}{\sqrt{1-x^2}} = \frac{\sqrt{2}}{\pi}\int_{-1}^{1-Nz}dx\frac{1}{\sqrt{1+x}}\left(1+\frac{z}{1-x}-\frac{z^2}{2(1-x)^2} +\frac{z^3}{2(1-x)^3}-\mathcal{O}\left(\frac{z^4}{(1-x)^4}\right)\right)\,. \tag{4.48}$$

The second integral can be simplified by noting that $1-x \ll 1$, and expanding the factor $\sqrt{1+x} = \sqrt{2-(1-x)}$ appearing in the denominator

$$\frac{\sqrt{2}}{\pi}\int_{1-Nz}^{1}dx\frac{\sqrt{1-x+2z}}{\sqrt{1-x^2}} = \frac{\sqrt{2}}{\pi}\frac{1}{\sqrt{2}}\int_{1-Nz}^{1}dx\frac{\sqrt{1-x+2z}}{\sqrt{1-x}}\left(1+\frac{(1-x)}{4}+3\frac{(1-x)^2}{32} +5\frac{(1-x)^3}{128}+\mathcal{O}\left(\frac{z^4}{(1-x)^4}\right)\right)\,. \tag{4.49}$$

Evaluating the integrals and adding (4.48) and (4.49) together, we get

$$\begin{aligned} a(z) &= \frac{1}{\pi}\left(4+z+\frac{3}{16}z^2-\frac{3}{32}z^3+\ldots\right)+(4\ln 2-\ln z)\left(z-\frac{z^2}{8}+\frac{3z^3}{64}-\frac{25z^4}{1024}\ldots\right) \\ &= \frac{1}{\pi}\left(4+z+\frac{3}{16}z^2-\frac{3}{32}z^3+\ldots\right)-ia_D(4\ln 2-\ln z) \quad \text{using (4.46)}\,. \end{aligned} \tag{4.50}$$

Writing everything in terms of u, the leading order is

$$a(u) \approx \frac{4}{\pi}-\frac{(u-1)}{2\pi}\ln(u-1)\,. \tag{4.51}$$

4.4.3 Solution in terms of Hypergeometric functions

An alternative, and concise expression for the integrals for $a(u)$ and $a_D(u)$ can be given in terms of hypergeometric functions $F(\alpha, \beta, \gamma; z)$. An integral representation of the hypergeometric function is given by (see, for example [28], [27])

$$F(\alpha, \beta, \gamma; z) = \frac{1}{B(\beta, \gamma - \beta)} \int_0^1 dt\, t^{\beta-1} (1-t)^{\gamma-\beta-1} (1-tz)^{-\alpha} \quad [\mathrm{Re}\,\gamma > \mathrm{Re}\,\beta > 0] \tag{4.52}$$

where

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt \tag{4.53}$$

is the beta function. We can recast the integrals for $a(u)$ and $a_D(u)$ in terms of $F(\alpha, \beta, \gamma; z)$ as follows. Recall that for the function $a_D(u)$, we have

$$a_D(u) = \frac{\sqrt{2}}{\pi} \int_1^u \frac{dx\,\sqrt{x-u}}{\sqrt{x^2-1}}\ . \tag{4.54}$$

Using the change of variable $x = (u-1)\,t + 1$ we get,

$$\begin{aligned} a_D(u) &= \frac{\sqrt{2}}{\pi} \int_0^1 \frac{dt\,(u-1)\,\sqrt{(u-1)t+1-u}}{\sqrt{(u-1)^2t^2 + 2(u-1)t}} \\ &= \frac{\sqrt{2}}{\pi} \int_0^1 \frac{dt\,(u-1)t^{-\frac{1}{2}}\,\sqrt{(1-u)\,(1-t)}}{\sqrt{u-1}\,\sqrt{(u-1)t+2}} \\ &= \frac{i\,(u-1)}{\pi} \int_0^1 dt\, t^{-\frac{1}{2}}\,(1-t)^{\frac{1}{2}} \left(1 - \left(\frac{1-u}{2}\right) t\right)^{\frac{-1}{2}} \\ &= \frac{i\,(u-1)}{2}\ \ F\left(\frac{1}{2}, \frac{1}{2}, 2; \frac{1-u}{2}\right). \end{aligned} \tag{4.55}$$

To simplify the function $a(u)$ starting from

$$a(u) = \frac{\sqrt{2}}{\pi} \int_{-1}^1 \frac{dx\,\sqrt{x-u}}{\sqrt{x^2-1}}, \tag{4.56}$$

we use the change of variable $x = 2\eta - 1$ to get,

$$\begin{aligned} a(u) &= \frac{\sqrt{2}}{\pi}\int_0^1 2\,d\eta\,\frac{\sqrt{2\,\eta - u - 1}}{\sqrt{4\eta^2 - 4\eta}} \\ &= \frac{\sqrt{2}}{\pi}\int_0^1 d\eta\,\frac{\sqrt{\left(\frac{2\eta}{1+u}\right) - 1}}{\sqrt{\eta(\eta-1)}}\,\sqrt{(1+u)} \\ &= \frac{\sqrt{2\,(1+u)}}{\pi}\int_0^1 d\eta\,\eta^{-\frac{1}{2}}\,(1-\eta)^{-\frac{1}{2}}\left(1 - \frac{2\eta}{1+u}\right)^{\frac{1}{2}} \\ &= \sqrt{2\,(1+u)}\;\; F\left(-\frac{1}{2}, \frac{1}{2}, 1; \frac{2}{1+u}\right). \end{aligned} \tag{4.57}$$

We can further simplify the expression for τ by using the complete elliptic integrals and their derivatives, along with the properties of hypergeometric functions. Consider the complete elliptic integrals of the first and second kind given in terms of the hypergeometric functions as

$$\begin{aligned} K(l) &= \frac{\pi}{2}F\left(\tfrac{1}{2}, \tfrac{1}{2}, 1; l^2\right), \\ E(l) &= \frac{\pi}{2}F\left(-\tfrac{1}{2}, \tfrac{1}{2}, 1; l^2\right). \end{aligned}$$

We immediately see that $a(u) = \frac{4}{\pi l}E(l)$, with $l^2 = \frac{2}{1+u}$.

4.4.4 Calculation of instanton numbers

Finally, we come to the computation of instanton numbers from the knowledge of the functions $a(u)$ and $a_D(u)$. Let us briefly recollect what we would like to compute. The full prepotential, with perturbative and non-perturbative corrections was found to be of the form, (3.4)

$$\mathcal{F} = \frac{i}{2\pi}a^2 \ln a^2 + \sum_{k=1}^{\infty} c_k\, a^2 \left(\frac{1}{a}\right)^{4k}. \tag{4.58}$$

The c_k terms represent the k-instanton corrections to the prepotential. We use an alternative expression of the functions $a(u)$ and $a_D(u)$ in terms of hypergeometric functions,

with different (α, β, γ) following the reference [26] where the instanton numbers were first computed

$$a_D(\alpha) = \frac{i}{4}(\alpha - 1)\ {}_2F\left(\frac{3}{4}, \frac{3}{4}, 2; 1-\alpha\right) \tag{4.59}$$

$$a(\alpha) = \frac{1}{\sqrt{2}}\alpha^{\frac{1}{4}}\ {}_2F\left(-\frac{1}{4}, \frac{1}{4}, 1; \frac{1}{\alpha}\right), \tag{4.60}$$

where $\alpha = u^2$. In the region where $u \sim \infty$ in the moduli space, the prepotential can be calculated as follows. First, we use the hypergeometric series representation

$$F(\rho, \mu, \sigma; z) = 1 + \frac{\rho \cdot \mu}{\sigma \cdot 1} z + \frac{\rho(\rho+1)\mu(\mu+1)}{\sigma(\sigma+1)\cdot 1 \cdot 2} z^2 + \frac{\rho(\rho+1)(\rho+2)\mu(\mu+1)(\mu+2)}{\sigma(\sigma+1)(\sigma+2)\cdot 1\cdot 2\cdot 3} z^3 + \dots \tag{4.61}$$

to express $a(u)$ with the help of a series in poweres of u using (4.60)

$$a(u) = \frac{\sqrt{u}}{\sqrt{2}}\left(1 - \frac{1}{16u^2} - \frac{15}{1024u^4} - \mathcal{O}\left(\frac{1}{u^6}\right)\right). \tag{4.62}$$

We want to invert this series for $a(u)$ to write a series for $u(a)$. To lowest order, we have

$$a \approx \frac{\sqrt{u}}{\sqrt{2}} \quad \Longrightarrow \quad u \approx 2a^2\ . \tag{4.63}$$

To find $u(a)$ to next order, we take $u = 2a^2 + \epsilon$, so that we have

$$\begin{aligned} a &= \frac{\sqrt{u}}{\sqrt{2}}\left(1 - \frac{1}{16u^2}\right) \\ &= \frac{\sqrt{2a^2+\epsilon}}{\sqrt{2}}\left(1 - \frac{1}{16u^2}\right) \\ &= a\sqrt{1 + \frac{\epsilon}{2a^2}}\left(1 - \frac{1}{16u^2}\right) \\ &\approx a\left(1 + \frac{\epsilon}{4a^2} - \frac{1}{16u^2}\right), \end{aligned} \tag{4.64}$$

giving us $\epsilon = \dfrac{1}{16a^2}$. To get the next order, we now take

$$u = 2a^2 + \frac{1}{16a^2} + 2\delta\,\frac{1}{a^6} = 2a^2\left(1 + \frac{1}{32a^4} + \delta\,\frac{1}{a^8}\right)\ . \tag{4.65}$$

Insetring this into the expression (4.62) for $a(u)$ retaining terms up to the order $1/a^8$, we get

$$\begin{aligned} a &= a\sqrt{1+\frac{1}{32a^4}+\delta\,\frac{1}{a^8}}\left(1-\frac{1}{64a^4(1+1/32a^4))^2}-\frac{15}{1024}\frac{1}{16a^8}\right) \\ &\approx a\left(1+\frac{(8192\,\delta-5)}{16384}\frac{1}{a^8}+\ldots\right), \end{aligned} \tag{4.66}$$

giving us $\delta = 5/8192$. Putting this back into (4.65), we get the expression for $u(a)$ to the next order to be

$$u(a) = 2a^2 + \frac{1}{16}\frac{1}{a^2} + \frac{5}{4096}\frac{1}{a^6} + \mathcal{O}(a^{-10}). \tag{4.67}$$

To obtain $\mathcal{F}(a)$, we need to obtain the expression for $a_D(u)$. Expanding the hypergeometric function ${}_2F\left(\frac{3}{4},\frac{3}{4},2;1-\alpha\right)$ we get from (4.59)

$$\begin{aligned} a_D(\alpha) &= \frac{i}{4}\,(\alpha-1)\ {}_2F\left(\frac{3}{4},\frac{3}{4},2;1-\alpha\right) \\ &= \frac{4}{\sqrt{2}\pi}\,\frac{i}{4}\alpha^{1/4}\left(\ln\alpha-4+6\ln 2-\frac{1}{16}\frac{\ln\alpha}{\alpha}+\frac{1}{\alpha}\left(\frac{1}{8}-\frac{3}{8}\ln 2\right)-\mathcal{O}\left(\frac{1}{\alpha^2}\right)\right)\ . \end{aligned}$$

Now, using $\alpha = u^2$, we get

$$a_D(u) = \frac{i\sqrt{u}}{\sqrt{2}\pi}\left(\ln u^2-4+6\ln 2-\frac{1}{16}\frac{\ln(u^2)}{u^2}+\frac{1}{u^2}\left(\frac{1}{8}-\frac{3}{8}\ln 2\right)+\mathcal{O}\left(\frac{1}{u^4}\right)\right)\ .$$

Using our result (4.67) for the expansion of $u(a)$, we get $a_D(a)$ as

$$a_D(a) \;=\; \frac{i}{\pi}a\ln\left(4a^4\right)+\frac{ia}{\pi}(-4+6\ln 2)+\frac{i}{\pi}\frac{1}{32a^3}+\frac{i}{\pi}\frac{15}{16384}\frac{1}{a^7}+\mathcal{O}\left(\frac{1}{a^{11}}\right). \tag{4.68}$$

Recall that a_D was defined as the derivative of the prepotential: $\partial\mathcal{F}/\partial a$. Thus, we can obtain $\mathcal{F}$ integrating (4.68):

$$\mathcal{F}(a) = \frac{ia^2}{2\pi}\left(\ln a^2-6+8\ln 2-\frac{1}{2^5}\frac{1}{a^4}-\frac{5}{2^{14}}\frac{1}{a^8}-\frac{3}{2^{18}}\frac{1}{a^{12}}+\frac{1469}{2^{31}}\frac{1}{a^{16}}+\mathcal{O}\left(\frac{1}{a^{20}}\right)\right)\ .$$

Thus, we can read off the first few instanton correction coefficients c_k to be

$$c_1 = \frac{1}{2\pi i}\frac{1}{2^5},\ c_2 = \frac{1}{2\pi i}\frac{5}{2^{14}},\ c_3 = \frac{1}{2\pi i}\frac{3}{2^{18}},\ c_4 = \frac{1}{2\pi i}\frac{1469}{2^{31}},\ldots\ .$$

We see that $\mathcal{F}$ is indeed of the expected form. Thus, we have computed the constants c_k corresponding to the k instanton corrections from the elliptic curve for $k = 1,\ldots,4$. This also gives us a tool to compute c_k to any order we wish, and hence determines $\mathcal{F}(a)$ completely. As discussed previously, this amounts to solving the low-energy effective theory.

4.5 Conclusion

In this thesis we have studied Riemann surfaces and one of their most important applications in theoretical physics - in the solution of Seiberg-Witten theory. We have studied the monodromy representation of the punctured u sphere, and shown that this corresponds to the monodromy associated to the modui space of $\mathcal{N} = 2$ supersymmetric SU(2) theory. We finally computed k-instanton contributions to the prepotential for $k = 1, \ldots, 4$.

References

[1] A. I. Bobenko, "Computational Approach to Riemann Surfaces," Lecture Notes in Mathematics, 2013.

[2] R. Miranda, "Algebraic Curves and Riemann Surfaces," Graduate Studies in Mathematics. Vol 5, 1997.

[3] F. Kirwan, "Complex Algebraic Curves," London Mathematical Society, Student Texts 23. 1999.

[4] B. Dubrovin, "Integrable Systems and Riemann Surfaces," Lecture Notes. 2009.

[5] D. A. Korotkin, "Introduction to functions on compact Riemann Surfaces and Theta-functions," arXiv:solv-int/9911002v1, 1999.

[6] M. Schlichenmaier, "An Introduction to Riemann Surfaces, Algebraic Curves and Moduli Spaces," Theoretical and Mathmatical Physics, 2007.

[7] M. Nakahara "Geometry, Topology and Physics," Institute of Physics Publishing, 2003.

[8] S. Donaldson, "Riemann Surfaces," Oxford Graduate Texts in Mathematics 22, 2011.

[9] H. M. Farkas, I. Kra "Riemann Surfaces," Graduate Texts in Mathematics. 1992.

[10] N. Seiberg and E. Witten, "Electro-Magnetic duality, monopole condensation, and confinement in $\mathcal{N} = 2$ supersymmetric Yang-Mills Theory", Nucl. Phys. **B426**, 1994. pp. 19–53, hep-th/9407087.

[11] N. Seiberg and E. Witten, "Monopoles, duality and chiral symmetry breaking in $\mathcal{N} = 2$ supersymmetric QCD", Nucl. Phys. **B431**, 1994. pp. 484–550, hep-th/9408099.

[12] M. Seiberg, "Supersymmetry and Nonperturbative beta functions," Phys. Lett. **B206**, pp. 75–80, 1988.

[13] N. Seiberg, "Exact results on the space of vacua of four-dimensional SUSY gauge theories", Phys. Rev. **D 49**, pp. 6857–6863, 1994.

[14] L. Alvarez-Gaume and S. F. Hassan, "Introduction to S duality in $\mathcal{N} = 2$ supersymmetric gauge theories: A pedagogical review of the work of Seiberg and Witten", Fortsch. Phys. **45**, 1997. pp. 159–236, hep-th/9701069.

[15] A. Bilal, "Introduction to Sypersymmetry", hep-th/0101055

[16] A. Bilal, "Duality in $\mathcal{N} = 2$ SUSY SU(2) Yang-Mills Theory: A pedagogical introduction to the work of Seiberg and Witten", hep-th/9601007

[17] E. D'Hoker and D. H. Phong, "Lectures on supersymmetric Yang-Mills theory and integrable systems", hep-th/9912271.

[18] S. V. Ketov, 'Solitons, Monopoles, and Duality", arXiv:hep-th/9611209v3, 1996.

[19] E. Witten and D. Olive, "Supersymmetry algebras that include topological charges," Phys. Lett. **78B**, pp. 97–101, 1978.

[20] M. F. Sohnius, "Introducing Supersymmetry", Phys. Rept. **128**, pp. 39–204, 1985.

[21] J. Terning, "Modern Supersymmetry," International Series of Monographs on Physics 132. 2006.

[22] P. Labelle, "Supersymmetry Demystified," McGraw-Hill Education 2010.

[23] H. J. W. M uller-Kirsten, and A. Wiedemann "Introduction to Supersymmetry," World Scientific Lecture Notes in Physics - Vol. 80.

[24] G. Springer, "Introduction to Riemann Surfaces," Addison-Wesley Publishing Company, 1957.

[25] R. Donagi, et.al "Integrable Systems and Quantum Groups," Lectures given at the 1st Session of the Centro Internazionale Matematico Estivo (C.I.M.E.) held in Montecatini Terme, Italy, June 14–22, 1993.

[26] W. Lerche, "Introduction to Seiberg-Witten theory and its stringy origin," Nucl. Phys. Proc. Suppl. **55B**, 1997. pp. 83–117, [Fortsch. Phys. **45**, 1997. pp. 293–340], hep-th/9611190.

[27] M. Abramowitz and I. Stegun, eds., "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables"' Dover Publications, 1972.

[28] I. S. Gradshteyn and I. M. Ryzhik, "Table of Integrals, Series, and Products Academic Press, 2007.